Universidad Politécnica de Madrid
Escuela Técnica Superior de Ingenieros de Telecomunicación

Contribución a las Metodologías de Estimación de Demanda de Tráfico de Internet mediante la Caracterización de Perfiles de Usuario

Tesis Doctoral

Mario Cao Cueto
Ingeniero de Telecomunicación

Año 2015

Departamento de Ingeniería de Sistemas Telemáticos
Escuela Técnica Superior de Ingenieros de Telecomunicación

Tesis Doctoral

Contribución a las Metodologías de Estimación de Demanda de Tráfico de Internet mediante la Caracterización de Perfiles de Usuario

Autor

Mario Cao Cueto

Ingeniero de Telecomunicación

Directores

Manuel Álvarez-Campana Fernández-Corredor

Doctor Ingeniero de Telecomunicación

Francisco González Vidal

Doctor Ingeniero de Telecomunicación

Año 2015

Tribunal nombrado por el Sr. Rector Magfco. de la Universidad Politécnica de Madrid, el día ..... de ..... de 2015.

Presidente:.....

Vocal:.....

Vocal:.....

Vocal:.....

Secretario:.....

Suplente:.....

Suplente:.....

Realizado el acto de defensa y lectura de Tesis el día ..... de ..... de 2015 en la E.T.S. de Ingenieros de Telecomunicación de Madrid.

Calificación:.....

EL PRESIDENTE LOS VOCALES

EL SECRETARIO

"Es ist nicht das Wissen, sondern das Lernen,
nicht das Besitzen, sondern das Erwerben,
nicht das Dasein, sondern das Hinkommen,
was den größten Genuß gewährt."

"No es el conocimiento, sino el aprendizaje,
no es la posesión, sino la adquisición,
no es el "estar allí", sino el acto de llegar,
lo que concede el mayor disfrute."

— Carl Friedrich Gauss

Agradecimientos

Llegado este momento, quiero agradecer de todo corazón a todas las personas que, con su trabajo, compañía y apoyo, me han acompañado a lo largo de esta etapa de mi vida y han contribuido, de forma directa o indirecta, a la consecución de esta tesis doctoral.

En primer lugar, quiero agradecer a mis directores de tesis, Manuel Álvarez-Campana y Francisco González Vidal, vuestra inestimable ayuda y apoyo durante todos estos años. Todo esto no hubiese sido posible sin esas palabras de ánimo que tantas veces me habéis regalado y todo el optimismo que me habéis transmitido. Gracias por haberme dado la oportunidad de aprender tantas cosas.

Me gustaría mostrar mi más sincera gratitud a todo el grupo de Redes y Servicios de Telecomunicación e Internet, especialmente a Julio Berrocal, Enrique Vázquez y Víctor Villagrá, por la confianza que habéis depositado en mí y todo el apoyo que me habéis brindado a lo largo de todos estos años.

También me gustaría extender mis agradecimientos al resto de miembros del Departamento de Ingeniería de Sistemas Telemáticos. En especial quiero agradecer a David Fernández, Luis Bellido y Encarna Pastor, el cariño y el apoyo desinteresado que siempre habéis sabido transmitirme.

Me gustaría mencionar de forma especial a todos los compañeros que habéis pasado por el laboratorio. A mis compañeros Alberto, Carlos, Joaquín, Jorge, Pedro, Pilar, Verónica y Vicente, muchas gracias por acompañarme en el día a día y por haber contribuido al excepcional ambiente de trabajo. A Jorge, gracias por ser el gran amigo que me has demostrado ser. A Carlos, gracias por tu apoyo y entusiasmo con el que nos contagias todos los días. A Verónica, gracias por estar siempre dispuesta a ayudarme. Y a Joaquín, empezamos esta aventura del doctorado prácticamente a la vez, gracias sobre todo por tu compañía y paciencia durante todos estos años que hemos trabajado juntos.

Además, me gustaría agradecer a la Asociación para la Investigación de Medios de Comunicación, y en especial a Fernando Santiago, por haberme recibido de forma tan agradable y haberme proporcionado desinteresadamente los datos estadísticos necesarios para el desarrollo de esta tesis doctoral.

De forma indirecta, este trabajo tampoco hubiese sido posible sin aquellos amigos,

con los que comparto otras parcelas de mi vida, y que me han demostrado su continuo cariño en forma de comprensión, paciencia y ánimos durante todo este tiempo. A vosotros, que sabéis quiénes sois, muchas gracias por todo.

Y por último, quiero destacar y dar las gracias desde lo más profundo de mi corazón a mi familia, a quienes realmente dedico este trabajo. A mis padres, a mi abuela Mercedes, a mi hermano Javier, a mis tíos Joaquín y Charo, y al resto de mi familia, muchas gracias por creer siempre en mí, por haberme apoyado y demostrado tanto cariño durante el prolongado, y en ocasiones difícil, periodo de tiempo que he dedicado a la realización de esta tesis doctoral.

Resumen

Esta tesis doctoral propone una metodología de estimación de demanda de tráfico de Internet basada en la caracterización de perfiles de usuario de Internet, con el objetivo de analizar el rendimiento y dimensionamiento de una red de acceso.

Se realiza un exhaustivo análisis del estado del arte clasificado en tres partes. La primera parte se encuentra relacionada con la caracterización de usuarios en Internet. Incluye un estudio de las metodologías de extracción de conocimiento basado en técnicas de minería de datos, y un análisis de modelos teóricos y estudios previos de usuarios de Internet. En la segunda parte, se incluye un análisis de modelos teóricos para caracterizar fuentes de tráfico de aplicaciones de Internet, así como un estudio comparativo de los modelos de tráfico ON/OFF para un conjunto de aplicaciones representativas de Internet. En la última parte, se incluye un estudio de las arquitecturas de redes de acceso más relevantes y se propone un modelo genérico de arquitectura de red de acceso.

Esta tesis doctoral define un marco metodológico basado en Procesos de Descubrimiento de Conocimiento (KDPs), con el que extraer, identificar y caracterizar, a los usuarios de Internet a partir de fuentes de información estadística. Se ha aplicado esta metodología a los usuarios residenciales en España y se ha identificado una distinción clara entre *No-Usuarios* (47 %) y *Usuarios de Internet* (53 %). Dentro de los usuarios de Internet se han extraído 4 perfiles de usuarios: *Esporádicos* (16 %), *Instrumentales* (10 %), *Sociales* (14 %) y *Avanzados* (13 %). Esta metodología también ha sido aplicada a años anteriores con el fin de realizar un pronóstico de la evolución de la tipología de usuarios de Internet en España.

A continuación, se propone un método de estimación de demanda de tráfico basado en los perfiles de usuario de Internet identificados, con el objetivo de analizar el rendimiento de la red de acceso subyacente. Esta metodología se encuentra basada en 3 modelos: red de acceso, tráfico de red y perfiles de usuario y aplicaciones.

Por último, la tesis presenta un modelo y una herramienta de simulación con la que se implementa el método de estimación de demanda anteriormente descrito. El modelo y la herramienta de simulación han sido validados frente a un modelo analítico mediante el uso de un escenario simplificado basado en fuentes de tráfico ON/OFF homogéneas.

Mediante el uso de la herramienta de simulación desarrollada, se aplica la metodología

de estimación de demanda a dos casos de uso, que se corresponden a dos escenarios de redes de acceso idénticas, a excepción de la caracterización de los usuarios de la misma. En el primer caso de uso, la red de acceso se caracteriza por los perfiles de usuario residenciales de Internet identificados para el año 2012, y en el segundo caso de uso, se utiliza el pronóstico de evolución de perfiles de usuario de Internet para el año 2017. Se concluye con una comparación del rendimiento de la red de acceso para ambos casos de uso, a partir del análisis del Grado de Servicio (GoS) de ambos escenarios.

Abstract

This thesis proposes a methodology for estimating the Internet traffic demand based on the characterization of user profiles in order to analyze the performance and dimensioning of access networks.

A comprehensive analysis about the state of art of 3 different knowledge areas is performed. The first knowledge area comprises the characterization of Internet users and the current methodologies for extracting typologies by using data mining techniques. It also includes an analysis of theoretical models and previous studies on Internet users. In the second knowledge area, this thesis performs a comparative study of traffic ON/OFF source models for a given set of representative Internet applications. In the last knowledge area, the most relevant access network architectures are described in order to propose a generic access network model.

First, this thesis proposes the use of a Knowledge Discovery Process (KDP) based methodology to identify and characterize Internet user typologies. Each step of the methodology is developed in detail and applied to a study case for Spanish Internet users, using a relevant and high quality data source. As a result, a clear distinction between *Non-Users* (47 %) and *Internet users* (53 %) is identified. Among the *Internet users* 4 user profiles have been extracted: *Sporadics* (16 %), *Instrumentals* (10 %), *Socials* (14 %) and *Advanced* (13 %). This methodology has also been applied to describe the evolution of the typology of Spanish Internet residential users in recent years and, based on this evolution, a prediction of the future evolution of the extracted Internet user profiles is presented.

Then, this thesis proposes a method to estimate the traffic demand based on the identified user profiles, which enables the performance and dimensioning analysis of the underlying access network. This methodology is based on three models: network access, network traffic, and user profiles and applications.

Finally, a simulation model is presented and a simulation tool is developed, which implements the aforementioned method for estimating the traffic demand. Both, model and simulation tool were validated against an analytical model by using a simplified scenario based on homogeneous traffic sources.

By using the developed simulation tool, the traffic demand estimation methodology

has been applied to two use cases. They represent two scenarios with the same characteristics excluding the Internet user characterization. In the first use case, the access network has been defined by means of the identified Internet typology of the year 2012. For the second use case, a forecast of the evolution of Internet user typology for the year 2017 is used. The access network performance is analyzed for both use cases in terms of the Grade of Service (GoS) provided by the network.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En los últimos años el tráfico de Internet se ha visto incrementado de forma notable y se espera que esta tendencia no cambie en los próximos años. Este incremento, con un factor cercano a cinco, ha sido posible gracias al despliegue de nuevas tecnologías de redes de acceso. No obstante, el motor de este cambio reside en los usuarios de Internet, pues sus hábitos y comportamientos de consumo de aplicaciones han ido evolucionando a lo largo de estos años y tienen un impacto directo en la demanda de tráfico de las redes de acceso. Por ejemplo, el consumo de aplicaciones de video sobre Internet ha experimentado un incremento significativo en los últimos años y se espera que siga aumentando en el futuro, lo cual ha producido que en el año 2013 cerca del 66 % de tráfico total se deba a este tipo de aplicaciones [Cisco, 2014].

En este contexto, la caracterización de usuarios de Internet mediante la identificación de su tipología, cobra una vital importancia a la hora de conocer el consumo de tráfico global de Internet. Esta tipología permite caracterizar a diferentes perfiles de usuarios a partir de sus comportamientos y hábitos de consumo de aplicaciones de Internet, lo cual puede utilizarse para estimar la demanda de tráfico que hacen de la red. Además, esta caracterización de perfiles de usuario puede utilizarse para estudiar la posible evolución en el tiempo de estas tipologías y para analizar el rendimiento e impacto de las demandas de tráfico de Internet en las redes de acceso actuales y futuras.

A continuación, se presentan los objetivos y la estructura de la memoria de esta tesis doctoral.

1.1. Objetivos

El principal objetivo de esta tesis doctoral es definir una metodología de estimación de demanda de tráfico de usuarios de Internet que permita analizar el rendimiento y el dimensionado de una red de acceso. La estimación de la demanda de usuarios de Internet estará basada en la caracterización de un conjunto de perfiles de usuarios y un

conjunto de aplicaciones representativas del tráfico de Internet. El análisis de la red de acceso y su dimensionado se apoyará en métricas de rendimiento asociadas al ancho de banda necesario en un enlace determinado.

Esta metodología de estimación de demanda de tráfico de Internet y dimensionado de red de acceso se aplicará a dos casos de estudio de redes de acceso: un escenario actual y un escenario que considere un pronóstico de evolución de los perfiles de usuarios de Internet en los próximos años. El objetivo de la aplicación de estos casos de estudio, es la de analizar el impacto en el rendimiento de la red de acceso a causa de un cambio en la tipología de usuarios de Internet.

En base a este objetivo global, se proponen los siguientes objetivos concretos:

- **Caracterización de usuarios de Internet en España** mediante una metodología específica que permita la extracción de una tipología de usuarios de Internet, que identificará un conjunto de perfiles de usuarios en función de sus hábitos, patrones de comportamiento y consumo de servicios de Internet a partir de datos obtenidos de fuentes de información estadística.

- **Propuesta de un método de estimación de demanda de tráfico y dimensionado de red de acceso** que tenga en cuenta la existencia de una tipología de usuarios y sus patrones de consumo de aplicaciones de Internet. Estimación que nos ha de servir para analizar el rendimiento que tiene una red de acceso.

- **Desarrollo de un modelo y herramienta de simulación** que implemente el mencionado método de estimación de demanda de tráfico y dimensionado de red de acceso.

- **Aplicación del método** de estimación de demanda de tráfico y dimensionado de red de acceso **a casos de estudio** representativos. Uno de los casos de estudio se basará en la tipología de usuarios de Internet extraída anteriormente, mientras que el otro caso de estudio, se basará en un pronóstico de tipología de usuario de Internet para los próximos años. Se analizará el impacto en el rendimiento que ha sufrido la red de acceso en término del Grade of Service (Grado de Servicio) (GoS) proporcionado.

1.2. Estructura de la memoria

Esta tesis doctoral se organiza en los siguientes capítulos:

- **Capítulo 1.** Introduce los objetivos, la motivación y la estructura de esta tesis doctoral.

- **Capítulo 2.** Se presenta el estado del arte de esta tesis, clasificado a su vez en tres secciones diferenciadas por el contexto al que hacen referencia:
  - Caracterización de usuarios de Internet: se presenta la metodología utilizada durante la caracterización de usuarios y se presentan un conjunto de estudios de la literatura y análisis de técnicas y métricas necesarias para las tareas de minería de datos.
  - Caracterización de tráfico de Internet: se describen algunos conceptos generales sobre distintos modelos de tráfico y se analizan multitud de modelos de tráfico ON/OFF para las aplicaciones de Internet más representativas.
  - Dimensionado de redes de acceso: se describen métricas de rendimiento de red y las arquitecturas de red de acceso más relevantes en la actualidad. Además, se propone un modelo genérico de red de acceso.
- **Capítulo 3.** Se desarrolla la caracterización de usuarios de Internet aplicando una metodología introducida en el capítulo anterior. Se describen con detalle todos los pasos llevados a cabo para extraer una tipología de usuarios de Internet a partir de una fuente de información estadística. Se incluye un análisis de calidad de los resultados obtenidos, una caracterización socio-demográfica de los perfiles de usuario de Internet identificados y un pronóstico de la tipología de usuario de Internet para los próximos años.
- **Capítulo 4.** Este capítulo comienza con la presentación del problema de acceso compartido a un recurso de red y la definición del GoS como una métrica de rendimiento de red de acceso. Posteriormente, se define una metodología para estimar la demanda de tráfico basada en 3 modelos: modelo de red de acceso, modelo de tráfico de red y modelo de demanda de tráfico de usuario.
- **Capítulo 5.** Se aplica la metodología anteriormente descrita para analizar el rendimiento de la red de acceso mediante la métrica de GoS para dos casos de estudio de red de acceso. Para la aplicación, se define un modelo de simulación y se desarrolla una herramienta de simulación, incluyendo una validación mediante el uso de un modelo analítico basado en fuentes de tráfico homogéneas.
- **Capítulo 6.** Se presentan las conclusiones más relevantes del desarrollo de esta tesis doctoral y se introducen las líneas futuras de investigación.

En la figura 1.1 se muestra el esquema de la metodología de estimación de demanda de tráfico y dimensionado de redes de acceso, propuesto en esta tesis doctoral. Se especifica para cada componente de la metodología, el capítulo de la tesis en la que se encuentra descrito. Los 3 componentes descritos en el primer nivel del esquema se

introducen y analizan en el capítulo 2 y 3. Estos componentes son utilizados para la definición del método para la estimación de demanda de tráfico en redes de acceso, descrito en el capítulo 4. En el capítulo 5, se describe la aplicación a dos casos de estudio, la cual permite visualizar mejor cómo encaja cada uno de estos componentes en el método de estimación de demanda de tráfico y dimensionado de red de acceso.

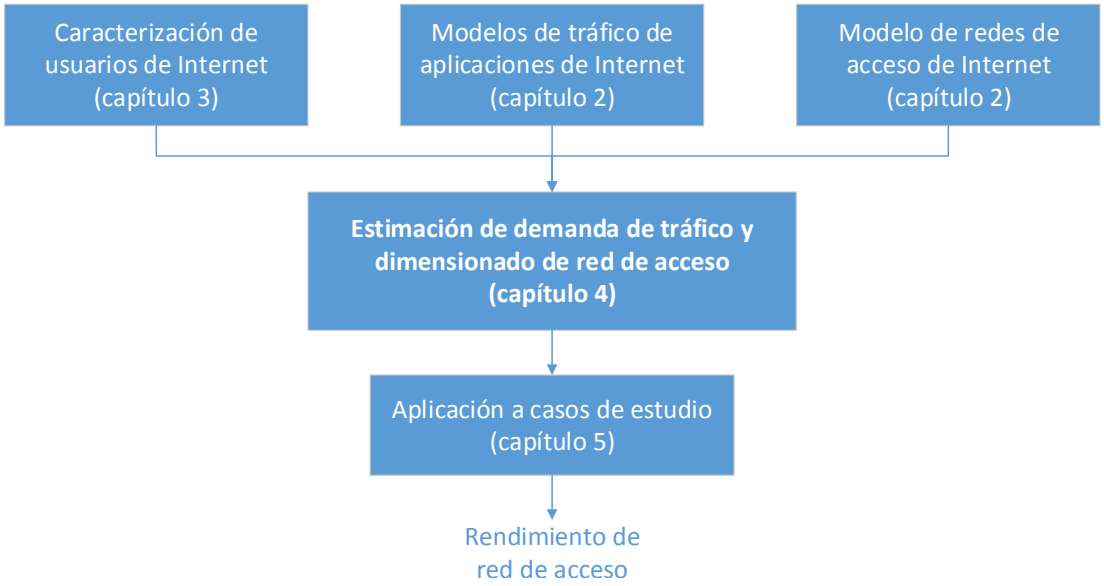

Figura 1.1: Esquema de metodología de estimación de demanda de tráfico y dimensionado de redes de acceso

Capítulo 2

Estado del arte

En este capítulo se aborda por separado el análisis del estado del arte de las tres áreas de conocimiento necesarias para la elaboración de esta tesis doctoral:

- Caracterización de usuarios de Internet
- Caracterización de tráfico de Internet
- Dimensionado de redes de acceso

El análisis de estas áreas de conocimiento aporta la base teórica necesaria para la definición de los componentes de entrada de la metodología de estimación de demanda de Internet, propuesta en esta tesis doctoral.

2.1. Caracterización de usuarios de Internet

En esta sección se aborda un análisis de la literatura de todas las referencias y herramientas necesarias para la caracterización de usuarios de Internet, descrita en el capítulo 3 de esta tesis doctoral.

2.1.1. Procesos de Descubrimiento de Conocimiento (KDP)

El método tradicional de convertir información en conocimiento se basa en el análisis manual e interpretación. Este análisis de los datos, además de requerir estar muy familiarizado con los mismos, es un proceso lento, costoso en tiempo y dinero, y de carácter muy subjetivo. Si además las fuentes de datos (por ejemplo, bases de datos) aumentan constantemente su tamaño y complejidad, nos encontramos ante un trabajo muy complejo y difícilmente realizable por un ser humano.

Ante esta necesidad de manejar cantidades de datos cada vez más grandes, surge la definición del Knowledge Discovery Process (Proceso de Descubrimiento de Conocimien-

to) (KDP), cuyo objetivo es identificar patrones y extraer información potencialmente útil y comprensible a partir de los datos [Fayyad et al., 1996b].

El KDP, también conocido como Knowledge Discovery in Databases (Descubrimiento de Conocimiento en Bases de Datos) (KDD), tiene como objetivo encontrar conocimiento nuevo a partir de un conjunto de datos en un dominio de aplicación determinado. Se define como un proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y comprensibles en un conjunto de datos [Cios et al., 2010]. A pesar de que en la literatura siempre se nombran a las bases de datos como fuentes principales de información, el KDP es de carácter general, por lo que también es válido para fuentes que no sean bases de datos. El KDP define por lo tanto una serie de pasos a seguir para poder extraer información de una fuente de información determinada.

Inicialmente los modelos de KDP aparecen en el mundo académico, aunque posteriormente la industria también comienza a realizar sus propios desarrollos. Las primeras propuestas datan del año 1996 de manos de [Fayyad et al., 1996a]. Este proceso consiste en una serie de pasos secuenciales, de forma que cada uno depende de los anteriores. Como se aprecia en la figura 2.1 también existen realimentaciones en los que los resultados de un paso pueden ocasionar la vuelta a un paso previo.

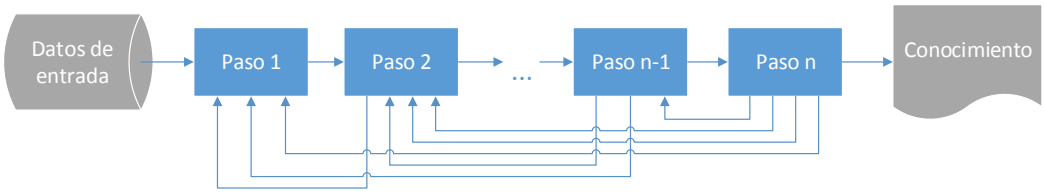

Figura 2.1: Estructura secuencial de los modelos de KDP

A continuación se describen los modelos de lo KDP más relevantes en la literatura, clasificándolos en 3 categorías: académicos, industriales e híbridos.

2.1.1.1. Modelos académicos

A mediados de los años noventa, comienza a ganar gran popularidad el campo de la minería de datos, por lo que se comienzan a definir algunos procedimientos que pretenden facilitar el descubrimiento y extracción de conocimiento. Estos modelos pretenden hacer especial énfasis en los pasos que se han de seguir para ejecutar con éxito un KDP en cualquier dominio.

El modelo más mencionado en la literatura se encuentra descrito en [Fayyad et al., 1996a] y que a su vez, se basa en el proceso definido en [Brachman and Anand, 1996]. A

continuación se enumeran los pasos definidos en el primero:

1. **Desarrollo y comprensión del dominio de aplicación.** Este paso hace referencia al conocimiento previo que se ha de tener en el dominio de aplicación, y la identificación de los objetivos del proceso KDD desde el punto de vista del usuario.

2. **Crear un conjunto de datos objetivo.** El usuario del proceso ha de seleccionar un conjunto de datos o focalizarse en un subconjunto de variables o muestras del mismo, a partir del cual extraerá el conocimiento.

3. **Limpieza de datos y preprocesamiento.** Las operaciones básicas en este paso son: la eliminación de ruido (si es necesario) y la definición de las estrategias a seguir con los valores perdidos.

4. **Reducción de datos y proyección.** Este paso consiste en encontrar los atributos de mayor utilidad para aplicar una reducción de dimensiones y métodos de transformación.

5. **Elección de la tarea de minería de datos.** En este paso el usuario ha de definir qué tarea se ha de realizar mediante la minería de datos. Esta tarea ha de ser coherente con los objetivos fijados en el primer paso del proceso.

6. **Elección del algoritmo de minería de datos.** Se elige el método de minería de datos para encontrar el conocimiento a partir de los datos y decidir que modelos y parámetros del método han de ser apropiados.

7. **Minería de Datos.** Este pase genera los patrones o la información extraída en una forma determinada de representación, por ejemplo, asociaciones, clasificaciones, árboles de decisión, modelos de regresión, etc.

8. **Interpretación de los resultados.** El analista realiza una interpretación de los resultados obtenidos, pudiendo retroceder a algunos de los pasos 1 hasta 7 para realizar nuevas iteraciones. Este paso también incluye la representación de los resultados mediante modelos de patrones o visualización de los modelos extraídos.

9. **Consolidación del conocimiento descubierto.** En este paso se utiliza los conocimiento adquiridos como entrada a otro sistema o simplemente mediante su correspondiente documentación e informando a terceras partes interesadas.

Un KDP exitoso puede realizar numerosas iteraciones entre los pasos anteriormente descritos. Además, a pesar de que el séptimo puede considerarse el más importante y más documentado en la literatura, el resto de pasos son esenciales para la correcta consecución del descubrimiento y extracción del conocimiento de un conjunto de datos determinado.

2.1.1.2. Modelos industriales

La industria también ha desarrollado y definido modelos de KDPs, entre los cuales destacan principalmente dos:

- Modelo propuesto en [Cabena et al., 1998] con el apoyo de IBM.
- Modelo CRoss-Industry Standard Process for Data Mining (Proceso Estándar industrial para la Minería de Datos) (CRISP-DM), desarrollado por un consorcio de compañías europeas y definido en [Shearer, 2000].

Es este último el que se ha terminado imponiendo en la industria como el modelo más utilizado y aplicado, ya que cuenta con el apoyo de grandes compañías y fue financiado por la Comisión Europea. En la figura 2.2 se ilustran los principales pasos del modelo y se aprecia cómo pueden existir realimentaciones entre los pasos que conforman el proceso, tal y como sucedía con los modelos académicos. El modelo define seis pasos diferentes para extraer el conocimiento de un conjunto de datos:

1. **Conocimiento del negocio.** Este paso inicial del proceso se centra en el entendimiento de los objetivos y requisitos desde el punto de vista del negocio, para poder definir posteriormente el problema de minería de datos y un plan preliminar para alcanzar los objetivos anteriores. Esta fase puede ser dividida a su vez en los siguientes pasos:

   a) Definición de objetivos de negocio
   b) Evaluación de la situación
   c) Definición de objetivos de la minería de datos
   d) Generación de un plan del proyecto

2. **Entendimiento de los datos.** Este paso comienza con la colección de datos iniciales y continúa con la familiarización con los mismos. De forma más específica, este paso intenta identificar problemas de calidad de los datos y detectar algunos subconjuntos de especial interés. Esta fase también puede ser dividida en los siguientes pasos:

   a) Colección de datos iniciales
   b) Descripción de datos
   c) Exploración de datos
   d) Verificación de calidad de datos

3. **Preparación de datos.** Este paso cubre todas las actividades necesarias para construir el conjunto de datos finales, sobre los cuales se van a aplicar métodos de minería de datos. Esta fase incluye los siguientes pasos:

   a) Selección de datos

   b) Limpieza de datos

   c) Construcción de datos

   d) Integración de datos

   e) Formateo de datos

4. **Modelado.** En este paso se seleccionan los métodos y técnicas adecuadas que se aplican. Esta fase puede llegar a requerir el uso de diferentes métodos de minería de datos para el mismo problema, ya sea para la correcta elección del método, o bien para la correcta calibración de parámetros. Esta fase se compone de los siguientes pasos:

   a) Selección de técnicas de minería de datos

   b) Generación de diseño de prueba

   c) Creación de modelos

   d) Evaluación de los modelos generados

5. **Evaluación.** Este paso tiene lugar cuando ya se han construido uno o varios modelos que ya han sido evaluados desde la perspectiva del análisis. Ahora bien, en esta fase se procede a una evaluación desde el punto de vista del objetivo del negocio. También se realiza una revisión de los pasos anteriores del proceso, con el objetivo de determinar si existen problemas que no hayan sido suficientemente considerados. Al final de la fase se decide qué pasos han de realizarse. Este paso se divide a su vez en:

   a) Evaluación de los resultados

   b) Revisión del proceso

   c) Determinación del próximo paso a realizar

6. **Despliegue.** Finalmente, el conocimiento descubierto ha de ser organizado y presentado de forma que el cliente o usuario pueda utilizarlo. En función de los requisitos, este paso puede ser tan sencillo como producir documentación, o tan complejo, como la implementación de un KDP que sea repetible. Esta fase se podría dividir en los siguientes pasos:

   a) Despliegue del plan

   b) Monitorización y mantenimiento del plan

   c) Generación de informe final

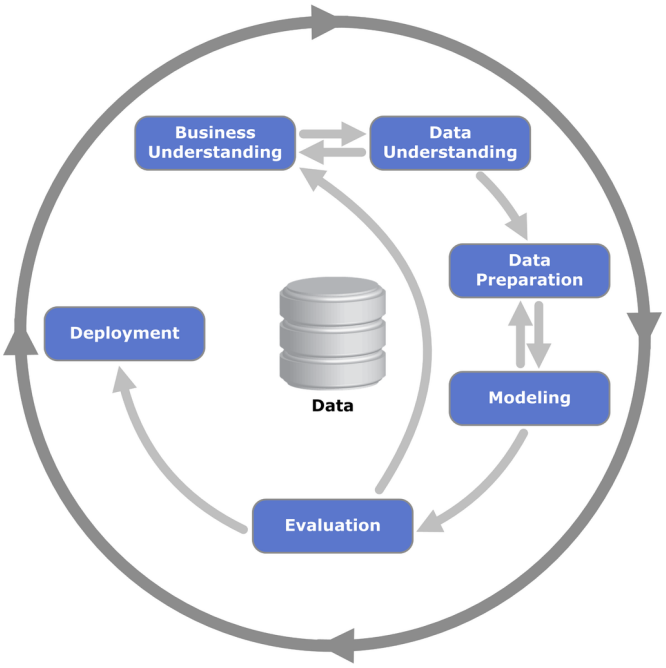

Figura 2.2: Diagrama del modelo CRISP-DM (Fuente: Wikipedia)

2.1.1.3. Modelo híbridos

Posterior a los modelos académicos e industriales, aparecen los modelos híbridos, entre los que destacan el propuesto por [Cios et al., 2010]. Este modelo híbrido se basa principalmente en el modelo CRISP-DM pero enfocándolo a la investigación académica. A continuación se describen los seis pasos en los que se compone el modelo:

1. **Comprensión del dominio del problema.** El paso inicial, al igual que en otros modelos, trata de comprender e identificar el dominio de aplicación y los objetivos del proyecto. Este paso también comprende la definición de objetivos y una selección inicial de la herramienta de la minería de datos a ser utilizada en las siguientes fases.

2. **Comprensión de los datos.** Este paso incluye la colección de las muestras de datos, y la decisión de qué formato y tamaño será necesario para el proceso de minería. Se realiza un chequeo de algunos parámetros de calidad de los datos, como por ejemplo, la completititud, redundancia, valores perdidos, etc.

3. **Preparación de los datos.** Este paso tiene como principal objetivo decidir qué datos van ser utilizados como parámetros de entrada para el método de minería de datos seleccionado. Esta fase incluye tareas de muestreo, tests de correlación, limpieza de datos, etc. Los datos que ya han sido limpiados, podrán

ser preprocesados por medio de algún algoritmo de extracción para reducir la dimensionalidad de los mismos. Al final de este paso, los datos resultantes han de satisfacer los requisitos definidos en el primer paso.

4. **Minería de datos.**. En este paso se procede a realizar la minería de datos a partir de los datos preprocesados en la fase anterior.

5. **Evaluación del conocimiento descubierto.** La evaluación incluye la comprensión de los resultados y el correspondiente de análisis en cuanto a la novedad y valor del conocimiento extraído. Además, se hace un revisión de los pasos anteriores del modelo con objeto de identificar aquellas acciones alternativas que podían haberse llevado a cabo. Finalmente, este paso concluye con la generación de una lista de errores.

6. **Uso del conocimiento descubierto.** Este paso final consiste en realizar una planificación de cómo se va a utilizar el conocimiento extraído con el modelo. Además, también se planifica la monitorización e implementación del conocimiento adquirido, y se genera la documentación del proyecto. Finalmente, el conocimiento se despliega donde aplique.

2.1.1.4. Comparación de modelos

La mayoría de los modelos KDP siguen una secuencia muy similar de pasos. Las fases más comunes y relevantes de los procesos descritos anteriormente son: comprensión del dominio, minería de datos y evaluación del conocimiento. No se puede afirmar que exista un modelo KDP mejor que otro, pues cada uno de ellos tiene sus puntos fuertes y débiles en función del dominio de aplicación y los objetivos que se persigan. Por este motivo, los modelos híbridos pueden constituir una buena elección ya que representan un compromiso entre los modelos académicos e industriales.

2.1.2. Modelos teóricos para la caracterización de usuarios

A continuación se presentan los tres enfoques de mayor relevancia asociados al ámbito de la caracterización de usuarios y con validez de aplicación para el caso particular de Internet.

2.1.2.1. Aceptación de tecnología

Los modelos basados en la aceptación de tecnología se originan a partir de la teoría de la acción razonada descrita en [Ajzen and Fishbein, 1977]. Este enfoque se considera uno de los más dominantes en relación con la comprensión y predicción de uso de

sistemas de información. Los factores que influyen en la decisión de cómo y cuándo se utiliza o adopta una tecnología son:

- **Utilidad percibida:** grado en el que una persona cree que usando un sistema va a mejorar su rendimiento en una tarea.
- **Facilidad de uso percibida:** grado en el que una persona cree que utilizar un sistema determinado no requiere un esfuerzo apreciable.

Dentro de este enfoque destaca el modelo propuesto en [Venkatesh et al., 2003], donde se realiza una revisión de los modelos más populares y se formula un nuevo modelo unificado de la aceptación de la tecnología por parte del usuario. Este enfoque se centra en las características que percibe el usuario y que afectan o influyen en la decisión de utilizar servicios, como por ejemplo, expectativas de rendimiento y de esfuerzo requerido, condiciones del entorno, actitud, etc.

En los últimos años se han definido algunas extensiones al modelo de aceptación de la tecnología, donde se define con más detalle aquellos parámetros que más influyen en el uso de las tecnologías. Por ejemplo, la teoría unificada de aceptación y uso de la tecnología propuesta en [Im et al., 2011], se establece que los factores que caracterizan el comportamiento de uso de la tecnología son los siguientes:

- Intención de comportamiento
  - Expectativas de rendimiento: grado de utilidad percibido que tendrá la adopción.
  - Expectativas de esfuerzo: grado de esfuerzo percibido que se necesita para adoptar la tecnología.
  - Influencia social: grado en el que una persona perciba la importancia que otros le dan a que adopte un método o sistema.
- Condiciones que facilitan: factores del entorno que hacen una acción fácil.

En [Venkatesh et al., 2012] se presenta otra extensión al modelo para el caso particular de consumo de tecnologías de la información. En este modelo se añaden varios factores a los anteriormente vistos:

- Motivación hedónica: grado de satisfacción o entretenimiento.
- Precio: coste económico asociado al uso de la tecnología.
- Hábito: costumbre de uso de adopción o uso tecnológico.

Además, esta extensión añade varias características individuales referentes al consumidor de tecnologías de la información que moderan los factores incluidos en el modelo. Estas características específicas de cada individuo son la edad, el género y la experiencia.

2.1.2.2. Difusión de la innovación

Este enfoque se encuentra íntimamente relacionado al modelo propuesto en [Rogers, 2010]. El modelo se centra en describir la adopción y uso de las innovaciones desde 4 puntos de vista diferentes:

- Proceso de difusión: analiza cómo se han difundido las nuevas ideas o innovaciones y a qué velocidad lo han hecho.
- Categorización de innovadores: en función de la tasa de adopción de la innovación, el modelo los clasifica como: innovadores, primeros seguidores, mayoría precoz, mayoría tardía, y rezagados.
- Proceso de decisión: se analizan cómo adoptan las personas las innovaciones y sus razones.
- Estudio de aceptabilidad: se identifican las características de las innovaciones que han influido, positivamente o negativamente, en la adopción de la misma por parte de los usuarios.

2.1.2.3. Usos y gratificaciones

Este enfoque se asocia por primera vez a las comunicaciones e investigación de medios en [Katz et al., 1999]. Los modelos basados en este enfoque no buscan conocer cómo se utiliza un medio, sino descubrir las gratificaciones que se esconden detrás y que influyen en el porqué del consumo. Estos modelos definen las gratificaciones asociadas al consumo de un medio de comunicación, y su correspondiente impacto en el uso (o intención de uso).

Dentro de este enfoque no se dan clasificaciones de usuarios, pero sí motivaciones (o gratificaciones) que pueden corresponderse a factores relevantes a tener en cuenta a la hora de caracterizar usuarios. En el modelo de [McQuail, 2010] se presentan las 4 motivaciones principales para el consumo de medios: información, entretenimiento, interacción social, e identidad personal. Otros estudios para otros usos, como por ejemplo [Höflich and Rössler, 2001] incluyen otras motivaciones para el uso como: disponibilidad, inmediatez e instrumentalidad.

2.1.3. Estudios previos sobre usuarios de Internet

Las segmentaciones de conjuntos de personas en función de sus comportamientos, preferencias o necesidades, son muy comunes en el ámbito de los estudios de mercado y la ingeniería de requisitos. Esta elaboración de categorías de tipos de usuarios también es conocida por los términos de tipologías o perfiles de usuario.

Existen varios estudios relevantes donde se describe el consumo de servicios y el comportamiento del usuario de Internet en una zona geográfica determinada. Algunos ejemplos se encuentran en los informes presentados por *Pew Internet Institute* para los Estados Unidos [Horrigan, 2007], por *OFCOM* para el Reino Unido [OFCOM, 2008], o incluso por la Organización para la Cooperación y Desarrollo Económicos (OECD) a nivel internacional [Montagnier and Wirthmann, 2011].

Las principales diferencias de estos estudios más relevantes encontrados en la literatura radican en los enfoques utilizados [Brandtzæg, 2010]:

- Enfoque global donde se describe el uso general de Internet sin entrar en detalles muy específicos relativos a los servicios y/o a los usuarios.
- Enfoque más específico sobre los servicios online de Internet utilizados por los usuarios (especialmente servicios relacionados con compras a través de la red)
- Enfoque más específico sobre las comunidades online y redes sociales.

A continuación se describen los estudios más relevantes encontrados en la literatura, teniendo especial consideración por aquellos que tienen un enfoque global sobre el uso de servicios de Internet.

Uno de los estudios más destacado por su relevancia, alcance y contexto temporal, se describe en [Brandtzæg et al., 2011]. A partir de los datos estadísticos de varios países europeos, se realiza un análisis de conglomerados para la identificación de 5 tipos de usuarios de Internet y que contribuyen a una mejor compresión de la segmentación digital:

- No-Usuarios (42 %): individuos que prácticamente nunca usan Internet.
- Usuarios esporádicos (18 %): usuarios caracterizados por un uso ocasional e infrecuente de Internet. Habitualmente, el servicio que más utilizan es el correo electrónico.
- Usuarios de entretenimiento (10 %): este grupo de usuarios tienen unos valores más altos en aquellos servicios relacionados con el entretenimiento, como por ejemplo, el visionado de televisión, descarga de juegos, música, chats, etc.
- Usuarios instrumentales (18 %): este tipo de usuarios realizan actividades orientadas a cumplir una meta, como por ejemplo, la búsqueda de información sobre bienes y servicios, la banca electrónica, comercio electrónico, etc.
- Usuarios avanzados (12 %): este grupo de usuarios tienen los valores más altos en cuanto a la intensidad y variedad de uso de servicios de Internet. A pesar de estar caracterizados por los valores más altos en casi todos los servicios, estos usuarios se encuentran orientados a actividades más instrumentales que de entretenimiento.

En [Brandtzæg, 2010] se define una tipología unificada basada en dos dimensiones: frecuencia de uso y variedad de los medios utilizados. Esta clasificación de usuarios tiene en cuenta las implicaciones sociales en uso de medios, otorgándole un papel especial a los servicios asociados a las redes sociales y comunidades en Internet. Los tipos de usuarios identificados en el estudio se muestran en la tabla 2.1, donde se indica el valor cualitativo asociado a la frecuencia y variedad de uso.

Tipo	**Frec.**	**Variedad**	**Actividad asociada**
No-Usuarios	Sin uso	Sin uso	No utilizan ningún servicio de la red
Esporádicos	Baja	Baja	Ninguna actividad en particular, interés bajo, recién llegados.
Discutidores	Media	Media	Discusiones y adquisición y compartición de información. Actividades con un propósito.
Entretenimiento	Media	Media	Juegos en red, visualización de videos. También usos avanzados, como generación de contenidos, programación y compras.
Socializadores	Media	Media	Socializar, mantener el contacto con amigos y familia. Vida social activa. Uso espontáneo y flexible.
Holgazanes	Media	Baja	Holgazanear y matar el tiempo.
Instrumentales	Media	Media	Orientados a la utilidad. Habitualmente actividades relacionadas con el trabajo. Poco uso de entretenimiento.
Avanzados	Alta	Alta	Todas las actividades. Gran variedad de servicios y uso intensivo.

Tabla 2.1: Tipología unificada basada en dos dimensiones [Brandtzæg, 2010]

En [Horrigan, 2009], *Pew Internet* se realiza un análisis sobre la vida digital de los usuarios de medios, haciendo especial hincapié en la movilidad de los mismos. El estudio identifica dos grupos principales: usuarios motivados por la movilidad (39 %) y usuarios estacionarios (61 %). Dentro de este segundo grupo se identifican además 5 segmentos de usuarios:

- Veteranos de escritorio (13 %): usuarios de mayor edad que principalmente exploran la red y mantienen el contacto con amigos. Las aplicaciones móviles permanecen en segundo plano.

- Surfistas a la deriva (14 %): usuarios con uso infrecuente de la red, principalmente para buscar información. A estos usuarios no les supondría un gran problema renunciar a Internet o a sus teléfonos móviles.

- Sobrecargados de información (10 %): estos usuarios sufren una sobrecarga de información y consideran que tomarse un tiempo apartado de la red es una buena idea.

- Indiferentes tecnológicos (10 %): usuarios poco interesados en la red, a pesar de disponer de teléfonos móviles. Tienen poco apego a los dispositivos tecnológicos y servicios de telecomunicación.

- Fuera de la red (14 %): las personas de este grupo no disponen de acceso o de teléfonos móviles, siendo habitualmente personas de mayor edad y con pocos ingresos económicos.

En [Horrigan, 2007] se realiza otro estudio de *Pew Internet* donde se clasifican a los usuarios de las Tecnologías de la Información y la Comunicación (TIC). Esta tipología de usuarios se realiza de forma exploratoria a partir de datos estadísticos, pero inspirada en la teoría de aceptación de la tecnología. Los segmentos de usuarios que se identifican son los siguientes:

- Usuarios tecnológicos de élite (31 %): uso intensivo y frecuente de las tecnologías de la información, incluyendo Internet y la telefonía móvil. Estos usuarios tienen un alto nivel de satisfacción en cuanto al rol que cumplen las TICs en sus vidas.

- Usuarios tecnológicos *en medio de la carretera* (20 %): entienden las TIC como un medio para cumplimentar tareas. Principalmente utilizan las TICs con fines comunicativos antes que como un medio para la autoexpresión.

- Usuarios con pocos activos tecnológicos (49 %): los dispositivos modernos se encuentran en la periferia de las vidas de este tipo de usuarios. Algunos encuentran estas tecnologías útiles, otros no y otros simplemente se mantienen en generaciones anteriores de móviles y televisores.

En [Ortega Egea et al., 2006] se presenta un estudio de gran calidad y relevancia en la literatura debido a la metodología empleada y la población sobre la que se realizó. En este análisis se identifican varios tipos de usuarios de Internet en Europa a partir de un análisis de conglomerados:

- Rezagados (16 %): uso ocasional e infrecuente de Internet. Los servicios rara vez se utilizan con fines privados.

- Confundidos y adversos (2 %): la frecuencia de uso de Internet es intermedia. Este tipo de usuarios muestra cierta confusión sobre los servicios de Internet, siendo una de sus características la alta variabilidad de uso en los mismos.

- Avanzados (16 %): uso frecuente de Internet no sólo para realizar tareas administrativas, sino también para otros propósitos. Estos usuarios realizan compras online con mayor frecuencia.

- Seguidores (19 %): uso de Internet con cierta frecuencia, pero no de forma diaria. Este grupo de usuarios no realizan compras online.
- No-Usuarios (44 %): El grupo con mayor número de integrantes lo conforman aquellos que nunca utilizan Internet.

En [Selwyn et al., 2005] se describe un estudio llevado en varias regiones de Inglaterra y Gales sobre el uso de Internet en la población adulta. Basándose en la tipología definida en [Howard et al., 2001], se realiza un análisis de frecuencias de datos estadísticos para identificar los siguientes grupos de usuarios:

- Usuarios frecuentes y con amplia variedad de uso (13 %): usan Internet de forma frecuente y con una variedad de 3 o más actividades.
- Usuarios frecuentes y con escasa variedad de uso (18 %): usan Internet de forma frecuente pero con una variedad de 1 o 2 actividades.
- Usuarios ocasionales (11 %): usan Internet de forma ocasional y/o esporádica.
- No-Usuarios (58 %): no han utilizado Internet durante los últimos 12 meses.

En [Shih and Venkatesh, 2004] se presenta un estudio realizado en Estados Unidos y basado en la teoría de la difusión de la tecnología, descrita anteriormente. Las variables contempladas para el análisis son la variedad de uso y las tasas de uso de servicios de Internet. Los tipos de usuarios identificados son los siguientes:

- Intensivos (30 %): usuarios con unos valores significativos en la tasa de uso (tiempo dedicado a la semana) y variedad de uso (número de aplicaciones utilizadas).
- Especializados (20 %): usuarios con una alta tasa de uso de servicios pero que sin embargo utilizan pocas aplicaciones.
- No-especializados (20 %): usuarios que tienen una tasa de uso de Internet baja pero una gran variedad de uso.
- Limitados (30 %): individuos caracterizados por una tasa y variedad de uso baja de servicios de Internet.

Uno de los primeros estudios que trataron de comprender el comportamiento y uso de Internet por parte de la población de Estados Unidos se describe en [Howard et al., 2001]. Este estudio se fundamenta en el modelo teórico de la difusión de la tecnología y realiza un análisis de frecuencias de uso desde los hogares para clasificar a los usuarios de Internet. A pesar de la antigüedad, este trabajo sigue siendo relevante en la literatura científica, ya que es utilizado a menudo como referencia bibliográfica a la hora de comparar los factores que caracterizan los perfiles de usuarios en Internet. Los tipos de usuarios identificados son los siguientes:

- *Netizens* (16 %): usuarios que llevan varios años conectándose y lo hacen de forma diaria. Internet forma parte de sus trabajos y vidas personales. No les importa gastar dinero online.

- *Utilitaristas* (28 %): usuarios experimentados en Internet y que también utilizan la red de forma diaria. A diferencia con los *Netizens*, estos usuarios tienen un uso menos intensivo y son menos activos en la red. Perciben Internet como una herramienta.

- Experimentadores (26 %): usuarios con una experiencia más limitada, pues conocen la red desde hace menos de 3 años. Utilizan Internet como una herramienta para buscar información.

- Recién llegados (30 %): tienen menos de un año de experiencia en Internet por lo que aún están aprendiendo a manejarse en la red. Realizan una gran variedad de actividades, muchas de ellas relacionadas con el entretenimiento, como por ejemplo, jugar en red, participar en chats, escuchar y bajar música, navegar, etc.

En [Johnson and Kulpa, 2007] se presenta una clasificación de las principales motivaciones que influyen en el uso de Internet. Se realiza a partir de encuestas en la población de Estados Unidos y se fundamenta en los tipos de personalidad y sociabilidad de los usuarios. Los factores identificados son coherentes con algunas de las características específicas de los tipos de usuarios encontrados en otros estudios:

- Sociabilidad: uso de Internet asociado a comportamientos sociales y con una motivación de conectar con otros individuos.

- Utilidad: uso de Internet típicamente instrumental y enfocado a la utilidad y eficiencia de alcanzar una meta.

- Reciprocidad: factor que describe el comportamiento online a partir de estímulos cognitivos y de participación activa de los individuos.

2.1.3.1. Estudios sobre usuarios de Internet en España

En el caso de España, existen muy pocas referencias bibliográficas que realicen una identificación de grupos o segmentos de usuarios de Internet. Un primer estudio a considerar, se encuentra en [Observatorio Nacional de las Telecomunicaciones y de la Sociedad de la Información (ONTSI), 2006], donde se presenta un estudio sociodemográfico sobre los usuarios de Internet en España y las actividades que realizan en la red. La principal variable de segmentación es la frecuencia de uso de Internet, a partir de la cual definen 6 grupos de usuarios de Internet entre la población española:

- Intensivos estables (32 %): usuarios con una frecuencia de uso alta y estable a lo largo de 24 meses.
- Incorporados (32 %): nuevos usuarios de Internet, pues hace 2 años no utilizaban la red.
- Progresivos (13 %): usuarios que han incrementado la frecuencia de uso de Internet hasta niveles próximos a los usuarios intensivos estables.
- Estables de uso medio (10 %): usuarios que mantienen una frecuencia parecida de uso a lo largo del tiempo y a niveles medios.
- Regresivos (8 %): usuarios que han disminuido su alta frecuencia de uso pero que se mantienen como internautas habituales o esporádicos.
- Ex Usuarios (4 %): individuos que han abandona Internet en los últimos 24 meses.

2.1.4. Fuentes de información

En esta sección se presentan las fuentes de información que disponen de datos de interés para la caracterización de usuarios de Internet, tanto de ámbito nacional como internacional.

2.1.4.1. Instituto Nacional de Estadística (INE)

El INE es un organismo autónomo de España encargado de coordinar los servicios estadísticos de la Administración General del Estado y la supervisión y de los procedimientos técnicos de los mismos. Además, el INE se encarga de las relaciones en materia estadística con los Organismos Internacionales especializados y, en particular, con la Statistical Office of the European Communities (Oficina Europea de Estadística) (EUROSTAT).

Una de las áreas de interés del INE son las nuevas tecnologías de la información y la comunicación, donde se encuentra una encuesta sobre el equipamiento y uso de tecnologías de la información y comunicación en los hogares españoles [INE, 2012]. Esta encuesta se elabora desde el año 2002 y tiene como objetivo recoger información sobre los diversos productos de las tecnologías de la información y comunicación de los hogares, así como del uso que le dan los usuarios de Internet a estos productos y a los servicios de la red. Además, esta encuesta sigue las recomendaciones metodológicas de la oficina estadística europea, EUROSTAT. En la tabla 2.2 se presenta la ficha técnica de la encuesta del año 2012.

Tabla 2.2: Ficha técnica: encuesta sobre el equipamiento y uso de tecnologías de la información y comunicación en los hogares españoles (INE)

Población:	Personas de 15 o más años que residen en viviendas familiares principales
Periodicidad:	Anual
Ámbito geográfico:	Todo el territorio español
Muestreo:	Muestreo trietápico estratificado
Criterio de estratificación:	Tamaño del municipio al que pertenece la sección
Método de recogida de información:	Encuesta personal y telefónica
Referencia temporal para el uso de Internet:	En los tres últimos meses
Número de muestras:	15.965 muestras
Disponibilidad de Microdatos:	Libre disposición desde el año 2002 al 2012

2.1.4.2. Asociación para la Investigación de Medios de Comunicación (AIMC)

La AIMC está formada por un amplio grupo de empresas cuya actividad gira en torno a la comunicación y cuyo principal interés es la de estudiar y conocer mejor todo lo relacionado con el consumo de medios en España. Entre las diferentes publicaciones generadas por esta asociación destacan principalmente dos estudios como posibles fuentes de información para la caracterización de usuarios de Internet.

El primer estudio analiza la audiencia de Internet [AIMC, 2013a] y tiene como objetivo primario, la estimación de la penetración de Internet entre la población española basándose en los datos recolectados en entrevistas personales dentro de otro estudio de ámbito más general llamado Estudio General de Medios (EGM). Este estudio cuenta con una metodología de recogida de información basada en entrevistas aleatorias de los hogares españoles, por lo que los resultados son representativos para todo el territorio nacional. Además, destaca la precisión de las preguntas de la encuesta donde se especifican diferentes servicios de Internet (jugar en red, ver videos, redes sociales, escuchar música, etc.). En la tabla 2.3 se presenta la ficha técnica correspondiente al EGM.

El segundo estudio, detallado en [AIMC, 2013b], describe los hábitos de internautas en cuanto a los servicios que utilizan en Internet. Se basa en encuestas disponibles en diferentes sitios web, por lo que los resultados pueden no representar a la totalidad de la población. La principal diferencia de este informe con el anterior, es que el EGM es representativo de la población española de 14 o más años. En la tabla 2.4 se detalla la ficha técnica del estudio.

Tabla 2.3: Ficha técnica: EGM (AIMC)

Universo:	Población española de 14 o más años (39.331.000 individuos)
Periodicidad:	Anual
Tamaño Muestral:	Muestra anual tres últimas olas: 30.844 entrevistas (última ola 10.744)
Muestra de la última ola:	10.744 entrevistas
Método de recogida de información:	Entrevista personal
Diseño Muestral:	Selección aleatoria de hogares y elección de una persona del hogar
Disponibilidad de Microdatos:	Sólo disponibles para socios de la AIMC desde el año 2000 al 2012

Tabla 2.4: Ficha técnica: Navegantes en la Red (AIMC)

Universo objetivo:	Los usuarios de Internet que visitan sitios web españoles
Periodicidad:	Anual
Tamaño Muestral:	35.213 cuestionarios (Muestra útil: 33.254)
Método de recogida de información:	Cuestionario a través de Internet a través de una aplicación desarrollada por la empresa ODEC
Fechas de recogida:	Encuesta activa en la red desde el 16 de Octubre hasta el 9 de Diciembre de 2012
Información adicional:	Los resultados fueron revisados y validados, eliminando aquellas entrevistas duplicadas y con algún tipo de irregularidad
Disponibilidad de Microdatos:	Libre disposición desde el año 2004 al 2012

2.1.4.3. Centro de Investigaciones Sociológicas (CIS)

El CIS es un organismo autónomo gubernamental y que tiene como principal objetivo el estudio científico de la sociedad española a partir de realización de encuestas periódicas.

Dentro de los estudios realizados por esta organización, destacan los barómetros mensuales que vienen realizando desde el año 1979, y donde se analizan principalmente aspectos sociodemográficos, económicos y políticos. No obstante, el CIS incluye en estos estudios mensuales, cuestiones relativas a temas de interés o de actualidad. Por esta razón, algunos de los barómetros incluyen información relativa a las tecnologías de la comunicación y la información, como por ejemplo, en los barómetros de mayo de 2013, junio y mayo de 2012, entre otros. Estos barómetros no permiten el análisis de los patrones de uso que hacen los españoles de los servicios de Internet. El barómetro de mayor relevancia para esta tesis doctoral es el realizado en junio del año 2012 [CIS, 2012], ya que presenta mayor número de cuestiones relacionadas con los servicios de

Internet. La Tabla 2.5 describe la ficha técnica del mencionado estudio.

Tabla 2.5: Ficha técnica: Barómetro de Junio 2012 (CIS)

Población:	Personas de 18 o más años de España
Periodicidad:	Mensual
Tamaño Muestral:	2.500 entrevistas
Método de recogida de información:	Entrevistas personales en domicilio
Muestras útiles:	2.482 entrevistas
Diseño Muestral:	Polietápico, estratificación proporcional por tamaño de hábitat
Disponibilidad de Microdatos:	Libre disposición desde el año 1979

2.1.4.4. Asociación de Empresas de Electrónica, Tecnologías de la Información, Telecomunicaciones y Contenidos Digitales (AMETIC)

AMETIC es una asociación empresarial que cuenta con más de 5.000 empresas asociadas y tiene como objetivo fomentar y promover el desarrollo del sector de la electrónica, las TIC, las telecomunicaciones y los contenidos digitales. Esta asociación tiene como una de sus principales actividades la elaboración de diferentes informes sobre la evolución del sector de las TIC, aunque habitualmente desde un punto de vista económico-financiero.

En [AMETIC, 2011] se presenta un informe anual de carácter principalmente económico, aunque incluye unas secciones dedicadas a analizar la evolución y situación actual de las TIC, los servicios de telecomunicación y el consumo de contenidos digitales, siempre desde un punto económico.

Posteriormente en [AMETIC, 2012a], AMETIC y la empresa *Accenture* realizan conjuntamente un informe sobre Internet en las redes móviles. Este estudio, basado en encuestas, caracteriza los servicios de Internet en dispositivos móviles y a los usuarios que los consumen. En la tabla 2.6 se describe la ficha técnica de los datos recogidos en este estudio. Los microdatos generados a partir de las encuestas no se encuentran disponibles para el público y tampoco pueden ser utilizados, ya que debido a su método de recogida de información, no son representativos de la totalidad de la población española.

Además, en [AMETIC, 2012b] se publica un nuevo informe donde se presenta la situación actual de distribución de contenidos multimedia en España desde un prisma económico, prestando especial atención a los ingresos de cada sector. El estudio realiza un análisis de los diferentes modelos de negocio para explotar, distribuir y proteger los contenidos, donde se presta especial atención a nuevas tendencias, como son por ejemplo los conceptos de *streaming* y de *free-to-play*. El informe también presenta una

Tabla 2.6: Ficha técnica: *Always On Always Connected* (AMETIC)

Población:	Personas de 14 o más años de 13 países diferentes
Periodicidad:	Puntual (año 2012)
Tamaño Muestral:	17.225 entrevistas
Método de recogida de información:	Encuestas online
Requisitos para el estudio:	Usuarios de Internet con dispositivo móvil con acceso a la Red móvil y residentes en España
Muestras útiles:	1.615 entrevistas
Diseño Muestral:	Selección aleatoria de hogares y elección de una persona del hogar
Disponibilidad de Microdatos:	No disponibles (datos recopilados por la empresa GFK)

comparativa de las diferentes plataformas de distribución de contenidos multimedia online disponibles en España.

2.1.4.5. Comisión del Mercado de las Telecomunicaciones (CMT)

La CMT es la Autoridad Nacional de Regulación (ANR) del sector de las telecomunicaciones en España. La CMT elabora un informe anual sobre el sector de las telecomunicaciones [CMT, 2011], que complementa los informes de la AMETIC para tener una visión del sector TIC en su conjunto. Este informe contiene un análisis detallado de la situación y evolución en el tiempo del sector, siempre desde un punto de vista económico. Se presta especial atención a las medidas de tipo regulatorio adoptadas en los últimos años. Destacar que las estadísticas están disponibles para ser consultadas de forma online. El informe presenta un análisis de los diferentes sectores de las telecomunicaciones: las comunicaciones fijas, las comunicaciones móviles, la banda ancha fija y los servicios audiovisuales. De gran utilidad son los datos aportados por la CMT en cuanto a cifras de líneas y operadores de comunicaciones en los territorios españoles. Este informe también recoge un análisis de la evolución en el tiempo y de las tendencias de la población en cuanto al uso de servicios audiovisuales.

2.1.4.6. ONTSI

El ONTSI es un órgano adscrito a la entidad pública empresarial *Red.es*, cuyo principal objetivo es el seguimiento y el análisis del sector de las telecomunicaciones y de la sociedad de la información. Entre los informes que generan periódicamente el ONTSI destaca el informe anual [ONTSI, 2013b], dónde se presenta un perfil sociodemográfico de los internautas a partir de los datos del INE. Destaca la caracterización del internauta en parámetros socio-económicos, ya que se analizan los perfiles en función del género,

edad, situación laboral, nivel de estudios, tamaño de hábitat y renta neta. Se presentan datos agregados sobre el uso que hacen la población de servicios de Internet análogos a los realizados en el anterior informe con datos del INE.

En [ONTSI, 2013a] se presenta un estudio que aborda la penetración de las TIC en hogares y empresas. Cuenta con la novedad metodológica, frente a los demás existentes en esta área, que construye indicadores a partir de los datos obtenidos en los muestreos de las facturas de los hogares encuestados. El informe resultante analiza la demanda en el segmento residencial y completa el conocimiento del sector adquirido por otros estudios e indicadores de otros organismos. La ficha técnica correspondiente a este estudio se encuentra en la Tabla 2.7.

Tabla 2.7: Ficha técnica: Las TIC en los hogares españoles. Datos de actitudes, usos, equipamiento y gasto TIC (ONTSI)

Población:	17.243.326 hogares Individuos de 15 y más años: 38,974 millones Individuos de 10 y más años: 41,141 millones
Tamaño Muestral:	3.131 hogares y 6.666 individuos 10+ años
Periodicidad:	Trimestral
Ámbito geográfico:	Península, Baleares y Canarias
Diseño Muestral:	Estratificación proporcional por tipo de hábitat, con cuotas de segmento social, número de personas en el hogar y presencia de niños menores de 16 años en el hogar
Método de recogida de información:	Encuestas
Referencia temporal para el uso de Internet:	En los tres últimos meses
Disponibilidad de Microdatos:	Datos desde Enero 2004 hasta Diciembre 2012 (No disponibles no disponibles al público)

2.1.4.7. Fundación Telefónica y Fundación Orange

Ambas fundaciones publican informes anuales que ofrecen una visión general acerca de la situación de España en la Sociedad de la Información, a partir de la recopilación de indicadores de diversas fuentes públicas y privadas.

La Fundación Teléfonica presenta en [Fundación Telefónica, 2013] su informe anual correspondiente al año 2012, donde se analiza la historia, evolución, desarrollo y adaptación de la Sociedad de la Información en España en la última década. Este informe analiza las principales tendencias de la Sociedad de la Información en España. Destaca el papel que están jugando la banda ancha móvil y los nuevos dispositivos móviles, como son los *smartphones* y las *tablets*, ya que España se encuentra en uno de los primeros

puestos en cuanto a penetración de estos dispositivos en el mundo. En referencia a la banda ancha fija también se está experimentando una subida significativa en las líneas de alta velocidad de acceso a Internet. En comparación con el año 2011, las líneas de entre 20 Mbps y 50 Mbps han aumentado un 111 % y las superiores a 50 Mbps un 60 %. Además, el número de líneas de fibra óptica hasta el hogar, Fiber To The Home (Fibra hasta el Hogar) (FTTH), se han doblado en el año 2012. Por último, este estudio comenta la gran importancia que está cobrando el acceso y consumo de contenidos digitales entre los usuarios de Internet.

La Fundación Orange publica un informe anual correspondiente al año 2012 en [Fundación Orange, 2012], donde presenta un análisis de la Sociedad de la Información en España, intentando prestar atención a todos los componentes que la conforman: servicios, contenidos, empresas y ciudadanos. Este estudio trata de analizar distintos indicadores del sector para hacer una cuidada comparación con el resto de países miembros de la Unión Europea.

Ambos informes tratan de sintetizar la información obtenida de diferentes fuentes de información, como son, por ejemplo, INE, AMETIC, EUROSTAT, etc.

2.1.4.8. EUROSTAT

La oficina estadística de la Comisión Europea, EUROSTAT cuenta con actividad estadística en áreas relacionadas con la industria, la ciencia y tecnología. Esta oficina cuenta con los datos de las organizaciones de estadística nacionales de cada país miembro de la Unión Europea, por lo que para el caso de España, los microdatos que disponen para el territorio español son los mismos que los que se pueden obtener a través del INE.

En [EUROSTAT and Seybert, 2012] se publica un informe donde se describe el uso de Internet en los hogares y por los individuos en 2012. Esta publicación se limita a recoger y presentar las estadísticas agregadas de los países miembros y compara algunas de ellas con la anualidad anterior. Sin llegar a hacer ningún tipo de caracterización de los usuarios de Internet, este informe concluye con la gran importancia adquirida por las redes móviles y los dispositivos móviles inteligentes. Además, recoge algunas nuevas tendencias como el uso de Internet por usuarios cada vez más jóvenes y su uso mientras éstos se desplazan.

2.1.4.9. International Telecommunication Union (Unión Internacional de Telecomunicaciones) (ITU)

La ITU publica desde hace varios años un informe anual [ITU, 2012a] donde se presenta un estudio sobre la Sociedad de la Información. Este documento tiene como objetivo extraer información de carácter económico relacionada con el desarrollo del sector de las TICs. Por un lado, se presentan datos sobre el número de abonados a los

diferentes servicios de acceso a las comunicaciones con datos similares a los que presenta la CMT en España pero a nivel internacional. Y por otro, se describe el desarrollo de las TICs mediante un índice que mide el acceso, la utilización y las capacidades de las infraestructuras.

En [ITU, 2012b] se presenta un estudio sobre el estado de la banda ancha en el año 2012, donde destaca la importancia de obtener conectividad a Internet de alta velocidad como un requisito esencial para la sociedad moderna. El informe revisa el estado actual y tendencias de la Sociedad de la Información en el mundo, y concluye rotundamente con que la conectividad a Internet es un factor clave para impulsar a los países en la futura economía digital.

Ambos informes contribuyen a la comprensión del estado del ancho de banda a nivel mundial y su impacto en las economías. No obstante, a pesar de que estos estudios tienen un marcado carácter económico y están enfocados a promover el desarrollo de las tecnologías en todo el mundo, no podrán ser utilizados directamente en la caracterización de usuarios de Internet.

2.1.5. Procedimientos asociados a la preparación de datos

En esta sección se describen algunos conceptos y técnicas relacionadas con uno de los pasos del KDP, la preparación de los datos.

2.1.5.1. Valores atípicos y perdidos

Una de las tareas comunes a realizar durante la fase de preparación de datos es la identificación y tratamiento de algunos valores excepcionales que se encuentran en la muestra de datos, como son los valores atípicos y perdidos.

La detección de valores atípicos es una parte indispensable del proceso de minería de datos. Dada una muestra, se considera un caso como un valor atípico, cuando se pueden identificar características que son significativamente diferentes a las del resto de casos de la muestra [Tan et al., 2006]. Esta tarea tiene asociada una complejidad elevada, pues hay que definir qué características y cuándo sus valores son significativamente diferentes a los del resto. Por esta razón, no existe ninguna definición matemática que sirva para la detección de todos los valores atípicos, ya que es considerada una tarea con una gran componente subjetiva y que, además, puede ocasionar falsas detecciones de valores atípicos que pueden introducir errores en la muestra.

En una colección de datos pueden existir valores perdidos, es decir, muestras con variables sin respuesta o en blanco. Estos valores pueden deberse por una mala recogida de la información o por el propio diseño del estudio. Los valores perdidos pueden ser tratados de forma similar a un valor atípico.

Ante la presencia de datos atípicos o perdidos, se pueden realizar diversas acciones ordenadas de menor a mayor en cuanto al impacto que tienen sobre los datos:

- Ignorar estos valores, ya que algunas técnicas de minería de datos son robustas frente a los mismos.
- Reemplazar estos valores por un valor determinada, como por ejemplo, un valor que se fija en función de otras variables o muestras.
- Filtrar aquellos casos con estos valores, con la consecuente reducción muestral de los datos.
- Invalidar la variable con estos valores, perdiendo por tanto la totalidad de la información contenida en las muestras de esta variable.

2.1.5.2. Técnicas de transformación de datos

La transformación de datos engloba cualquier proceso en los que se modifica la forma de los datos. Estos procesos consisten, por tanto, en definir nuevos conjuntos de variables que son producto de transformaciones de los datos, derivaciones o cambios de tipo o rango.

Existen muchas posibles transformaciones que se pueden aplicar a los datos, pero entre las más comunes se encuentran las siguientes: discretización de variables, numerización de variables y normalización de variables.

Discretización. La discretización de datos, también conocida como *binning*, se basa en la conversión de datos numéricos en valores nominales, habitualmente ordenados, es decir, en variables ordinales o de tipo *likert* [Likert, 1932]. Este proceso suele realizarse ante la presencia de datos con bastantes errores o cuando existen umbrales significativos en las variables (por ejemplo, umbral en las notas de una asignatura para indicar el aprobado). Otro uso de la discretización se da cuando los rangos en los que se encuentran los datos son más relevantes que los propios valores numéricos (interpretación no lineal). También en función de la técnica de minería de datos, puede darse el caso en el que el algoritmo sólo soporte variables nominales. La discretización más sencilla se denomina *simple binning* y consiste en utilizar intervalos del mismo tamaño, utilizando los mínimos y máximos como referencias. Otra técnica sencilla se basa en obtener intervalos con el mismo número de registros o frecuencias de variables, conocido como *equal-frecuency binning*).

Numerización. La numerización de variables es el proceso inverso al anterior, en el cual se convierten variables nominales u ordinales en variables de tipo escalar. El uso de

esta técnica es muy común para poder aplicar ciertos algoritmos del proceso de minería de datos que sólo soportan atributos numéricos, como por ejemplo, aquellos basados en distancias. Existen dos tipos de numerizaciones de variables:

- Numerización *1 a n*: se crean tantas variables numéricas como posibles valores tiene una variable nominal dada. Los valores serán 0 o 1 dependiendo si la variable nominal toma el valor que correspondiente a la variable escalar. Estas variables también se conocen como variables de pertenencia, porque el 1 significa que se cumple cierta característica, y el 0 lo contrario.

- Numerización *1 a 1*: si existe cierto orden en la variable nominal se pueden numerar los valores correspondientes siguiendo este mismo orden, convirtiendo la nueva variable en una variable ordinal.

Normalización. Por último, la normalización de variables es un proceso mediante el cual se adecuan los datos para su correcto procesamiento posterior o poder ser más fácilmente interpretables. Este proceso suele ser aconsejable antes de utilizar técnicas de minería de datos basadas en distancias

2.1.6. Técnicas de minería de datos

Durante la elaboración de esta tesis doctoral se han utilizado diversas técnicas para realizar la minería de datos. Las técnicas utilizadas pueden ser clasificadas en función de la utilidad que proporcionan para la extracción del conocimiento:

- Técnicas descriptivas y de reducción de dimensionalidad

- Técnicas de análisis de conglomerados cuya finalidad es dividir un conjunto de objetos en diferentes grupos de forma que los objetos de un mismo grupo sean muy similares entre sí, y los objetos de grupos diferentes muy distintos.

2.1.6.1. Técnicas descriptivas y reducción de dimensionalidad

A continuación, se describen brevemente algunas de las técnicas utilizadas en esta tesis doctoral con las siguientes dos finalidades:

- Comprender la información que se encuentra en los datos

- Contribuir con este conocimiento a una posible reducción de la dimensionalidad de los datos

Análisis factorial. El análisis factorial tiene como objetivo describir la variabilidad observada de un conjunto de variables correladas mediante unas variables no observadas, denominadas factores o variables latentes. Además, este método puede ser utilizado para la reducción de la dimensionalidad del conjunto de datos, mediante la identificación de aquellos factores de menor relevancia.

Partiendo de la hipótesis de que las variables observadas $x' = (x_1, x_2, \ldots, x_p)$ no se encuentran incorreladas entre sí, se asume un número de factores $f_1, f_2, \ldots, f_k$, donde $k < p$ y de forma que se cumple la ecuación 2.1.

$$\begin{aligned} x_1 &= \lambda_{11} \cdot f_1 + \ldots + \lambda_{1k} \cdot f_k + u_1 \\ &\vdots \\ x_p &= \lambda_{p1} \cdot f_1 + \ldots + \lambda_{pk} \cdot f_k + u_p \end{aligned} \tag{2.1}$$

Los valores λ_{ij} son los pesos factoriales que muestran como cada x_i depende de variables latentes comunes. Para valores altos de λij relacionan íntimamente un factor con la variable observada, de forma que se posibilita una interpretación de cada uno de los factores.

Los valores $u_1, \ldots . u_p$, se denominan variables específicas y se asume que se encuentran incorrelados entre sí y con los factores $f_1, f_2, \ldots, f_k$.

Debido a que los factores no son observables, se puede considerar que son variables estandarizadas (media 0 y varianza 1) y se encuentran incorreladas entre sí. De este modo, los pesos factoriales λ_{ij} son las correlaciones entre variables y factores. Así pues, la varianza de x_i sigue la ecuación 2.2 donde Ψ_i es la varianza de u_i. Se puede observar como la varianza de cada variable se puede descomponer en dos partes: la comunalidad (h_i^2) y la varianza específica (ψ_i).

$$\sigma_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2 + \Psi_i = h_i^2 + \psi_i \tag{2.2}$$

En la ecuación 2.3 se muestra la covarianza entre dos variables distintas, donde se aprecia que ésta sólo depende de los factores comunes y no de las variables específicas.

$$\sigma_{ij} = Cov(x_i, x_j) = Cov(\sum_{l=1}^{k} \lambda_{il} \cdot f_l, \sum_{l=1}^{k} \lambda_{jl} \cdot f_l) = \sum_{l=1}^{k} \lambda_{il} \cdot \lambda_{lj} \tag{2.3}$$

A partir de la ecuación anterior, se puede obtener una matriz de covarianzas Σ para cada una de las variables observadas (ecuación 2.4), siendo Ψ la matriz diagonal cuyos componentes son las varianzas específicas. El análisis de esta matriz indica que existen variables altamente intercorrelacionadas entre sí, de forma que el análisis factorial tenga sentido y validez. En caso contrario, si las correlaciones entre las variables son bajas, el análisis factorial puede no ser apropiado. Para conocer el grado de asociación entre

variables existen diferentes indicadores como, por ejemplo, el test de esfericidad de Barlett [Dziuban and Shirkey, 1974].

$$\Sigma = \Lambda\Lambda' + \Psi \tag{2.4}$$

En la práctica, se estiman los parámetros a partir de una muestra de los mismos, por lo que el problema reside en encontrar valores de $\hat{\Lambda}$ y $\hat{\Psi}$ tal que (ecuación 2.5):

$$S \approx \hat{\Lambda}\hat{\Lambda}' + \hat{\Psi} \tag{2.5}$$

En un análisis factorial no existe una única solución pues existen diferentes métodos de extracción de factores, los cuales ofrecen resultados distintos. Entre los métodos de extracción de factores existentes, destacan el método de los factores principales y el método de máxima verosimilitud.

El método de los factores principales, es una técnica basada en autovalores y autovectores, y en la matriz de covarianzas reducida, que se calcula siguiendo la ecuación 2.6 y en las comunalidades estimadas de las variables observables.

$$S* = S - \hat{\Psi} \tag{2.6}$$

El método de la máxima verosimilitud, se basa en asumir la normalidad de los datos y en el uso de una métrica de distancia F entre la matriz de covarianzas observada y los valores predichos de esta matriz, por el modelo de análisis factorial (ecuación 2.7). La estimación de los pesos del análisis factorial se obtienen minimizando esta función, lo cual es lo mismo que maximizar la función de verosimilitud del modelo del análisis factorial.

$$F = ln\left|\Lambda\Lambda' + \Psi\right| + traza(S\left|\Lambda\Lambda' + \Psi\right|^{-1}) - ln\left|S\right| - p \tag{2.7}$$

Por último, destacar la importancia de la elección de un número adecuado de factores k para representar las covarianzas observadas. En función del número de factores se pueden encontrar pesos factoriales muy distintos.

Una vez extraída la matriz con los pesos factoriales, se requiere interpretar los resultados. En la práctica, estos métodos de extracción de factores pueden proporcionar matrices de cargas factoriales poco adecuadas para su interpretación. Para facilitar esta interpretación, existen procedimientos de rotación de factores que, a partir de una solución inicial del análisis factorial, proporcionen matrices de cargas factoriales más fácilmente interpretables. En función del tipo de rotación (ortogonal u oblicua) existen diferentes métodos, como por ejemplo, *varimax*, *oblimin*, *quatimax*, etc. Estas rotaciones no alteran la comunalidad de las variables ni la bondad del ajuste de la solución proporcionada por el análisis factorial. Estos procedimientos cambian la varianza

explicada por cada factor con objeto de simplificar la interpretabilidad de los resultados.

Análisis de componentes principales. El punto de partida de un análisis de componentes principales es el mismo que el visto anteriormente en el análisis factorial. De hecho, este análisis es en realidad un análisis factorial que debido a su simplicidad y aplicabilidad ha adquirido una entidad estadística propia.

El objetivo del análisis de componentes principales es la transformación de un conjunto de variables observadas en un nuevo conjunto de variables incorrelacionadas ente sí, denominadas componentes principales. Mediante este proceso no sólo se puede conseguir una reducción de la dimensionalidad de los datos, sino que también se puede describir la cantidad de información que contienen.

Partiendo de un conjunto de variables originales $x_1, x_2, \ldots, x_p$ se calcula un nuevo conjunto de variables $y_1, y_2, \ldots, y_p$, donde cada componente y_j es una combinación lineal de las variables originales (ecuación 2.8), siendo a'_j un vector de constantes.

$$y_j = a_{j1}x_1 + a_{j2}x_2 + \ldots + a_{jp}x_p = a'_j x \tag{2.8}$$

La idea tras el análisis de componentes principales es la de maximizar la varianza de las variables mediante un proceso iterativo. El primer paso del análisis de componentes principales es la de calcular el autovector a_1 de modo que y_1 tenga la mayor varianza posible y sujeto a la restricción de $a'_1 a_1 = 1$ (ortogonalidad de la transformación). El segundo componente principal se calcula obteniendo a_2 de modo que la variable y_2 esté incorrelada con y_1. Este proceso se itera hasta encontrar todas las variables $y_1, y_2, \ldots, y_p$, de forma que se encuentren incorreladas (o sean ortogonales) entre sí y que la variables obtenidas tengan cada vez menor varianza.

El análisis de componentes principales puede ser utilizado también como un método de extracción de factores para el análisis factorial mediante la selección de los dos o tres primeros componentes principales.

Reducción de la dimensionalidad. Cuando se realiza un experimento (por ejemplo, un experimento físico o simplemente una encuesta a una gran cantidad de individuos) se tiende a sobredimensionar los datos que se quieren recopilar para evitar posibles faltas de información que requerirían repetir el experimento por completo y, por tanto, un coste económico que puede ser significativo. Además, hay que añadir el hecho de que no todos los datos presentes en las fuentes de información fueron recolectados con el objetivo de ser utilizados como parámetros de entrada de alguna técnica de minería de datos. A causa de esto, las fuentes de información tienden a tener multitud de variables que hacen muy difícil las tareas de extracción de conocimiento a partir de KDP si no se reducen las dimensiones de los datos a ser utilizados como entradas del proceso.

Las razones por las que es recomendable aplicar la reducción de la dimensionalidad son varias, de entre las que destacan especialmente las dos siguientes:

- La maldición de la dimensionalidad: este fenómeno se introduce por primera vez en [Bellman, 1956], donde se afirma que la mayoría de los métodos y algoritmos de minería de datos no son muy efectivos cuando los datos de entrada presentan un gran número de dimensiones.
- Las dimensiones intrínsecas para definir un modelo pueden ser pocas: el análisis de conglomerados puede realizarse con un subconjunto de la muestra disponible.

Las técnicas más comunes para reducir las dimensiones de una colección de datos suelen ser las siguientes:

- Selección de características: se busca un subconjunto de las características definidas en variables que sean adecuadas para la extracción de un modelo.
- Extracción de características: se utiliza un algoritmo que a partir de las características, busca un nuevo conjunto de variables, con menor dimensionalidad.

En esta tesis doctoral se hace uso de la selección de características para encontrar un subconjunto adecuado para la extracción del conocimiento del KDP, la cual puede realizarse siguiendo dos enfoques diferentes:

- Métodos de filtro o métodos previos: se filtran las variables que no son relevantes con el KDP.
- Métodos basados en modelo o métodos de envolvente (*wrapper*): se evalúa la bondad de la selección de variables respecto a la calidad del modelo extraído a partir de los datos.

Esta técnicas se basan en realizar numerosas iteraciones donde se van añadiendo y/o eliminando variables. En cada una de ellas se analizan los resultados obtenidos a partir de la técnica de minería de datos seleccionada y se busca maximizar la calidad del modelo extraído. De acuerdo con la estrategia de búsqueda de subconjuntos de variables se puede establecer la siguiente taxonomía:

- Completa: se cubren todas las combinaciones posibles de selección de variables
- Heurística: se reduce el número de combinaciones a evaluar basándose en algún tipo de información.
- No determinista (estocástico): basada en algoritmos de búsqueda globales que intentan evitar el problema de mínimos locales.

En cuanto a la dirección de la búsqueda también se puede realizar la siguiente clasificación:

- Hacia adelante: se comienza con la mejor variable y se le añade otra variable con la que se obtenga mayor calidad con dos atributos. Este proceso se repite hasta que no se mejore la calidad o se llegue al número deseado de dimensiones.
- Hacia atrás: el proceso comienza con todos los atributos y se van eliminando uno a uno los atributos menos relevantes. De forma análoga, el proceso se repite hasta conseguir el número de dimensiones deseado o no se obtenga mayor calidad.
- Aleatoria: se producen patrones de búsqueda mediante la generación de conjuntos aleatorios.

Por último, destacar la complejidad asociada a la selección de un subconjunto adecuado, debido a que las combinaciones de variables crece exponencialmente con el número de características consideradas para el KDP (ecuación 2.9).

$$N_{comb} = \sum_{k=1}^{n} \binom{n}{k} = \sum_{k=1}^{n} \frac{n!}{k!(n-k)!} \tag{2.9}$$

2.1.6.2. Técnicas de análisis de conglomerados

En esta sección se presentan las técnicas de minería de datos más relevantes para la consecución de esta tesis doctoral, es decir, teniendo presente que el objetivo consiste en caracterizar a los usuarios de Internet y extraer una tipología asociada al uso que realizan de los servicios de Internet.

Las técnicas de agrupamiento o conglomerados pueden ser de dos tipos [Tan et al., 2006]:

- Supervisado: se clasifican los casos en base a unas categorías predefinidas, denominadas clases.
- No-Supervisado: no existe categorías predefinidas y la técnica consiste en descubrir o extraer una clasificación a partir de los datos, en función de la similitud de los casos entre sí.

A su vez, existen multitud de algoritmos de análisis de conglomerados en la literatura, pudiendo ser clasificados en función del modelo de agrupamiento que éstos siguen [Han et al., 2006]:

- Agrupamiento jerárquico: los algoritmos de este tipo se basan en la idea principal de relacionar los casos con aquellos más similares. El resultado es una jerarquía de casos que se representa en forma de dendograma o grafos (árboles).

- Agrupamiento particional: el objetivo de este tipo de algoritmos es el de categorizar o dividir los casos de la muestra en varios grupos, llamados particiones.

Las técnicas de agrupamiento jerárquicas son recomendables para análisis que requieran una gran granularidad de detalles. No obstante, este tipo de algoritmos adolecen de una baja eficiencia debido a que necesitan calcular tablas de distancias entre todos los casos. En caso de disponer de muestras de cierto tamaño, el agrupamiento jerárquico es inviable.

El agrupamiento particional es preferible para conjuntos de datos de tamaño considerable debido a que los algoritmos de este tipo tienden a tener una eficiencia mucho más alta que los basados en agrupamiento jerárquico.

Algoritmos basados en agrupamiento particional. Dentro de las técnicas basadas en agrupamiento particional existen numerosos algoritmos que se pueden clasificar en función del modelo de conglomerados que sigan. A continuación se presentan los 3 tipos de algoritmos basados en agrupamiento particional que gozan de mayor relevancia y popularidad en la literatura:

- Basado en centroides: se define un conglomerado como un conjunto de casos, los cuales se encuentran más cerca (o son más similares) al centroide del segmento que a los centroides de los otros. El centroide del conglomerado es la media de todos los puntos que conforman un segmento. Habitualmente, este tipo de algoritmos requieren que se especifique previamente el número de conglomerados a buscar. Dentro de esta familia de algoritmos, destaca el algoritmo K-Means [Lloyd, 1982].

- Basados en distribuciones estadísticas: este tipo de algoritmos agrupa los casos en diferentes segmentos teniendo en cuenta criterios estadísticos de los mismos. Los casos se van agrupando en función de las probabilidades calculadas de pertenecer al segmento. Uno de los algoritmos de este tipo más utilizado, a pesar de su antigüedad, es el algoritmo EM (Expectation Maximization) [Dempster et al., 1977].

- Basados en densidades: se definen un conglomerado como una región de alta densidad de casos y que se encuentra separada de otros conglomerados por regiones de baja densidad de casos. Las zonas de baja densidad de casos están compuestas normalmente por ruido y valores atípicos de la muestra. Uno de los algoritmos más populares y del cual se basan muchos otros, es el algoritmo DBSCAN [Ester et al., 1996].

No existe un algoritmo mejor que otro, sino más adecuado para el problema que se quiere resolver. Los algoritmos basado en centroides cuentan con una buena escalabilidad y eficiencia al analizar grandes cantidades de datos. Este hecho permite realizar

numerosos análisis para la comparación y validación de resultados. Por contrapartida, este algoritmo tiene como requisito que se ha de conocer el número de conglomerados a extraer

Los algoritmos basados en distribuciones estadísticas, como el algoritmo EM, tienen escalabilidad más limitada y la complejidad asociada al algoritmo aumenta considerablemente con el número de parámetros de entrada.

En el caso de los algoritmos basados en densidad, es importante destacar que no tienen en cuenta los casos que se encuentran en zonas de baja densidad. Para colecciones de datos con una alta dispersión en los casos, considerará muchos de los casos como ruido o valores atípicos, no siendo incluidos en los resultados.

Algoritmo K-Means. En esta tesis se hace uso del algoritmo *K-Means*, el cual se clasifica como algoritmo de agrupamiento particional basado en centroides. Este algoritmo requiere 3 parámetros que han de ser especificados previamente por el usuario: número de clusters K, centroides iniciales y una métrica de distancia, o similitud.

K-Means es un algoritmo sencillo e iterativo que puede descomponerse en los siguientes pasos:

1. Definición de los centroides iniciales. Los centroides iniciales pueden ser un parámetro de entrada o pueden ser definidos mediante el uso de diferentes estrategias.

2. Asignar cada muestra al centroide más cercano mediante el uso de una métrica de distancia o similitud.

3. Recalcular los valores de los centroides de los clusters como la media de los valores de las muestras que pertenecen a ese cluster.

4. Repetir los pasos 2 y 3 de forma iterativa hasta que las muestras no cambien de cluster o se llegue a un máximo de iteractiones.

A continuación (pseudocódigo 1), se muestra el pseudocódigo de una implementación básica del algoritmo *K-Means*, dónde se siguen los pasos anteriormente descritos.

A pesar de que este algoritmo es ampliamente utilizado en multitud de aplicaciones científicas, cuenta una serie de desventajas intrínsecas [Peña et al., 1999], entre las que destacan las siguientes:

- Se asume que se conoce el número de conglomerados K que se quieren extraer, lo cual no siempre se cumple, pues en gran cantidad de aplicaciones se desconoce este parámetro.

- Debido a que *K-Means* es un algoritmo iterativo, éste es especialmente sensible a las condiciones iniciales del mismo.

Pseudocódigo 1 K-Means: algoritmo básico

Entrada: $E = e_1, e_2, ..., e_n$ (conjunto de muestras para ser conglomeradas)
$C = c_1, c_2, ..., c_k$ (centroides)
$MaxIters$ (máximo de iteraciones)
Salida: $C = c_1, c_2, ..., c_k$ (conjunto de centroides de clusters)
$L = l(e)|e = 1, 2, ..., n$ (etiquetas de pertenencia a cluster de E)

```
1:  foreach e_i ∈ E do
2:    l(e_i) ← argminDistance(e_i, c_j), j ∈ 1, ..., k;
3:  end
4:  changed ← false;
5:  iter ← 0;
6:  repeat
7:    foreach c_i ∈ C do
8:      UpdateCluster(c_i);
9:    end
10:   foreach e_i ∈ E do
11:     minDist ← argminDistance(e_i, c_j), j ∈ 1, ..., k;
12:     if minDist ≠ l(e_i) then
13:       l(e_i) ← minDist;
14:       changed ← true;
15:     end
16:   end
17:   iter + +;
18: until changed = true ∧ iter ≤ MaxIters
```

- Este algoritmo puede converger a mínimos locales, lo cual se traduce en soluciones que no tienen por qué ser cercanas a la solución óptima o ideal.

A pesar de que no existe un criterio matemático exacto para la correcta selección del número de clusters que se han de extraer, existen numerosos enfoques heurísticos que abordan este problema [Tibshirani et al., 2001]. La gran mayoría de enfoques se basan en la repetición del algoritmo con diferentes valores de K para posteriormente seleccionar aquel que tenga una mayor calidad según una métrica de calidad determinada.

En [Milligan, 1980] se evidencia la dependencia existente entre la calidad de los resultados de análisis de conglomerados respecto a los valores iniciales de los centroides utilizados. Este hecho junto a la convergencia del algoritmo a mínimos locales, puede ocasionar la extracción de modelos sub-óptimos. Este problema se puede abordar mediante el uso de métodos de selección de condiciones iniciales del algoritmo que no sean puramente aleatorios. En [Bradley and Fayyad, 1998] se describen mecanismos de selección de centroides iniciales basados en las muestras de datos que van a ser utilizadas. De la misma forma, las herramientas de estadística, como por ejemplo *SPSS* de *IBM*, también ofrece un método de selección de condiciones iniciales basada en el cálculo de centros bien separados formados por casos de la muestra [SPSS, 2011].

Otro parámetro a seleccionar para el algoritmo es la métrica de similitud, la cual es considerada fundamental en el análisis de conglomerados. En algunos casos, el éxito del análisis de conglomerados puede depender más de la selección de una métrica de similitud apropiada que de la selección de un buen algoritmo [Hastie et al., 2005]. K-Means habitualmente se encuentra asociada a la distancia euclidea [MacQueen et al., 1967], aunque existen numerosas aplicaciones del algoritmo donde se utilizan otras métricas.

La métrica de Minkowski es utilizada en muchas versiones del algoritmo [Jain, 2010]. Esta métrica define la distancia ente dos puntos $P = (x_1, x_2, \ldots, x_n)$ y $Q = (y_1, y_2, \ldots, y_n)$ según la ecuación 2.10.

$$d = (\sum_{i=1}^{n} |x_i - y_i|)^{1/p} \tag{2.10}$$

Como se puede apreciar, distancia euclidea es un caso particular de esta métrica ($p = 2$). Las métricas de Minkowski pueden ocasionar cierta susceptibilidad a que algunas características prevalezcan sobre otras, para lo cual se recomienda que se normalicen las variables para que éstas tengan un mismo rango [Mao and Jain, 1996].

2.1.7. Métricas de calidad de conglomerados

Existe un gran número de métricas con el objetivo de cuantificar la calidad y la validez en función de la similitud entre los conglomerados extraídos a partir de técnicas de minería de datos. Estas métricas se pueden clasificar en función del criterio considerado en la evaluación de la calidad del método de extracción de segmentos:

- Evaluación externa: se analizan las distancias de los conglomerados extraídos frente a referencias ya dadas, denominadas etiquetas de clase. Estas referencias son datos que no han sido utilizados en el proceso de conglomerados y que ya han sido pre-clasificados en diferentes tipos.

- Evaluación interna: se analizan las características intrínsecas del análisis de conglomerados. Habitualmente estas métricas dan mejor puntuación a aquellos resultados donde existe una gran similitud dentro de un mismo conglomerado, y una baja similitud entre conglomerados.

La evaluación externa sólo es posible cuando existe un subconjunto de datos ya pre-clasificados. De esta forma se pueden contrastar las características de los conglomerados extraídos con las características de los casos pre-clasificados. En el caso de un análisis de conglomerados no-supervisado y donde no se ha realizado este proceso de pre-clasificación, este tipo de evaluación no es posible.

La evaluación interna suele ser el criterio utilizado cuando no se dispone de etiquetas de clase de un subconjunto de los datos. No obstante, aunque estas métricas premien la similitud dentro del conglomerado y las diferencias entre conglomerados, no siempre una buena puntuación es sinónimo de una buena efectividad o validez del modelo extraído en un dominio de aplicación determinado. Por esta razón, es recomendable que una evaluación basada en un criterio interno, se encuentre siempre acompañada de una evaluación directa frente al modelo conceptual que se esperaba extraer.

Las métricas más relevantes en la literatura y de especial interés para esta tesis doctoral, han sido seleccionadas por los siguientes tres criterios: popularidad de la medida, éxito de la medida en otros estudios, y simplicidad de implementación y coste computacional.

Los criterios que predominan en la gran mayoría de métricas de calidad son [Baarsch and Celebi, 2012]:

- Cohesión de los conglomerados: mide el grado de relación entre objetos dentro de un conglomerado.
- Separación de los conglomerados: mide el cómo de distintos o cómo de separados son los conglomerados entre sí.

Además, en algunos casos es muy importante que medida de evaluación tenga en cuenta el número de conglomerados que se extraen. Un modelo con un número de perfiles alto puede cumplir con los criterios anteriormente mencionados, pero conceptualmente puede no estar abstrayendo perfiles de usuario válidos.

En [Milligan and Cooper, 1985] se realizó un estudio comparativo de más de 30 métodos de evaluación, entre los que destaca la métrica definida por Calinski y Harabasz [Caliński and Harabasz, 1974]. Posteriormente, otros autores han propuesto muchas otras métricas de evaluación, entre las que despunta la métrica *Gap Statistic* descrita en [Tibshirani et al., 2001]. Este artículo describe la nueva métrica a la vez que la compara con las más relevantes y populares de la literatura. Debido a la baja complejidad, sencillez y buenos resultados de evaluación obtenidos en estudios comparativos [Milligan and Cooper, 1985, Tibshirani et al., 2001, Baarsch and Celebi, 2012], la métrica de Calinski y Harabasz se presenta como una buena alternativa a la métrica *Gap Statistic* anteriormente citada.

2.1.7.1. Métrica de Calinski-Harabasz (CH)

La métrica de Calinski y Harabasz $CH(k)$ se basa en la relación entre la matriz de dispersión entre conglomerados, Between Cluster Scatter Matrix (BCSM), y la matriz de dispersión dentro de los conglomerados, Within Cluster Scatter Matrix (WCSM). La BCSM es la suma de los cuadrados de las distancias del centro de cada conglomerado

(C_i) con el centro del conjunto de datos ($\bar{x}$), ponderado por el tamaño del conglomerado ($\frac{1}{N}$) (ecuación 2.11). La WCSM es la suma de los cuadrados de las distancias entre el centro del conglomerado (C_j) con cada punto (x_j) que pertenece al mismo (2.12). La métrica de Calinski-Harabasz se define según la ecuación 2.13, donde se aprecia que su puntuación disminuye a medida que el número de conglomerados (k) en la solución crece.

$$BCSM_k = \frac{1}{N}\sum_{i=1}^{k}|C_i - \bar{x}| \tag{2.11}$$

$$WCSM_k = \frac{1}{N}\sum_{i=1}^{k}\sum_{j\in C_i}|x_j - C_j| \tag{2.12}$$

$$CH(k) = \frac{BCSM_k}{WCSM_k}\frac{N-k}{k-1} \tag{2.13}$$

2.1.7.2. Métrica de Zhao (BW)

En [Zhao et al., 2009] se presenta otro métrica de calidad $WB(k)$, también basado en sumas cuadráticas, que realiza una comparación con otras métricas popularmente utilizadas en la literatura, como por ejemplo: Dunn's index (DI), Davies-Bouldin's Index (DBI), Bayes Information Criterion (BIC) y Silhouette Coefficient (SC). El índice $WB(k)$ dota de gran peso a la WCSM ya que le aplica un factor igual al número de conglomerados (ecuación 2.14).

$$BW(k) = \frac{WCSM_k}{BCSM_k}\cdot k \tag{2.14}$$

2.2. Caracterización de tráfico de Internet

En esta sección se presenta un análisis de los diferentes enfoques existentes para la caracterización de tráfico de Internet. Los objetivos principales son los siguientes:

- Revisión del estado del arte sobre los diferentes enfoques para la caracterización de tráfico de Internet.
- Definición de un conjunto representativo de servicios de Internet.
- Análisis y selección de modelos de tráfico del conjunto representativo de servicios de Internet.

Este análisis tiene como principal objetivo la selección de un conjunto de modelos de tráfico de un conjunto de servicios de Internet. Estos modelos de tráfico serán utilizados en la metodología de estimación de demanda en redes de acceso de esta tesis doctoral.

2.2.1. Motivaciones

Un área muy importante de investigación hoy en día, es el desarrollo de modelos de tráfico de red, que puedan ser aplicados a Internet o a cualquier otra red de comunicación.

Los modelos de tráfico pueden ser utilizados para la realización de simulaciones de red, las cuales intentan extraer resultados que de forma analítica resultaría muy costoso o prácticamente imposible debido a la complejidad de las redes de comunicaciones representadas. Estas simulaciones pueden tener infinidad de objetivos, como por ejemplo:

- la validación de algoritmos y protocolos,
- el análisis de tráfico ante diferentes condiciones de red.

Es de vital importancia que los modelos se acerquen en la medida de lo posible a las características del tráfico real que representan.

Estos modelos de tráfico también pueden ser utilizados como base para el diseño y planificación de capacidad de redes de comunicaciones, debido a la importancia del rendimiento ofrecido por las mismas. La gestión del rendimiento de las redes incluye la optimización de algunas de sus características, como puede ser maximizar la capacidad, minimizar la latencia y disponer de una alta confiabilidad. En definitiva, los modelos de tráfico permiten realizar suposiciones sobre las redes que están siendo diseñadas, basándose en experiencias pasadas y habilitando predicciones de rendimiento para futuras necesidades.

Las principales motivaciones para el desarrollo de modelos de tráfico de red son:

- Dimensionado y planificación de rendimiento de sistemas de red
- Gestión de rendimiento de la red
- Garantías de Quality of Service (Calidad de Servicio) (QoS)

Los modelos de tráfico se utilizan para dimensionar adecuadamente los recursos de red para un nivel determinado de rendimiento, es decir, de QoS. Estos modelos se utilizan para la estimación de anchos de banda, retardos y pérdidas de paquetes aceptables en la red. Un buen conocimiento de algunos parámetros de red, como por ejemplo la variabilidad de tráfico o *burstiness*, es requisito para determinar las capacidades de los enlaces a lo largo del sistema [Barakat et al., 2003].

El término de QoS hace referencia al nivel de rendimiento ofrecido por una red de comunicaciones a las aplicaciones y/o usuarios de la misma. La QoS se mide cuantitativamente mediante los análisis de diferentes parámetros de la red, tales como la tasa de errores, el ancho de banda, el retardo, la disponibilidad, etc. Algunos ejemplos de mecanismos de QoS son la priorización de tráfico y la garantía de un ancho de banda mínimo.

2.2.2. Niveles de actividad de tráfico

El análisis del tráfico proporciona información, como por ejemplo, la carga media, los requisitos de ancho de banda de diferentes aplicaciones y muchas otras utilizadas como parte de un modelo analítico o bien para ser utilizadas en Discrete Event Simulation (Simulación de Eventos Discretos) (DES).

El tráfico de red consta de un conjunto de llegadas de entidades discretas, es decir, paquetes, celdas, sesiones, llamadas, etc. Habitualmente, desde el punto de vista matemático, se utilizan dos representaciones fundamentales para modelar el tráfico:

- procesos que representan la aparición de un evento en el sistema
- procesos que representan el tiempo entre llegadas

Los parámetros más relevantes generados por la gran mayoría de modelos de tráfico son: distribuciones de longitud de datos (en términos de paquetes, ráfagas de paquetes, flujos, etc.) y distribuciones de tiempo entre llegadas.

En concordancia con el modelo Transmission Control Protocol (TCP)/Internet Protocol (IP) de Internet, la figura 2.3 muestra los diferentes niveles en los que el comportamiento de usuario y de aplicaciones puede ser modelada y medida [Grimm and Schlüchtermann, 2008]. Se asume que en cada nivel se puede determinar el principio y fin de eventos, así como el volumen de datos transferidos. Aunque habitualmente los modelos se encuentran definidos para un nivel de actividad determinado, existen algunos modelos que tienen en cuenta las características de diferentes niveles.

- **Nivel de acceso.** Este nivel hace referencia al tiempo en el que un usuario accede al sistema o red de comunicaciones, como por ejemplo, Internet. En el caso de usuarios con conexión permanente a la red, este tiempo puede corresponderse al tiempo de encendido de los equipos de conexión de los usuarios, como por ejemplo, un router.

- **Nivel de aplicación.** En este nivel se representan las aplicaciones, teniendo en cuenta que varias de éstas pueden estar siendo ejecutadas de forma paralela. De forma similar al nivel anteriormente descrito, los eventos se caracterizan principalmente por los periodos de actividad e inactividad de las aplicaciones.

- **Nivel de diálogo.** El siguiente nivel hace referencia a las interacciones entre usuarios y aplicaciones, las cuales generan flujos de datos a través de la red. Para ilustrar este nivel se pone como ejemplo las tres fases en las que se puede descomponer los diálogos de usuarios web: hacer *click* inicia el diálogo, esperar a que se descarguen los contenidos de una página web y el tiempo de lectura del contenido descargado.

- **Nivel de conexión.** Este nivel representa el tiempo entre conexiones realizadas entre entidades a más bajo nivel desde el punto de vista del modelo de referencia OSI. Continuando con los ejemplos anteriores, este nivel hace referencia a las conexiones del protocolo TCP, es decir, desde el establecimiento de conexión (*three way handshake*) hasta que finaliza. En el caso de que una aplicación utilice un protocolo de transporte no orientado a conexión, como por ejemplo User Datagram Protocol (UDP), no se podría calcular los tiempos correspondientes a este nivel.

- **Nivel de ráfaga.** Este nivel tiene en cuenta las ráfagas de paquetes, como es el caso de las ráfagas de paquetes IP, que contienen, a su vez, segmentos TCP o UDP.

- **Nivel de paquete.** Este es el nivel de menor abstracción considerado, en el que se tiene en cuenta los paquetes individuales a nivel de red, es decir, a nivel IP.

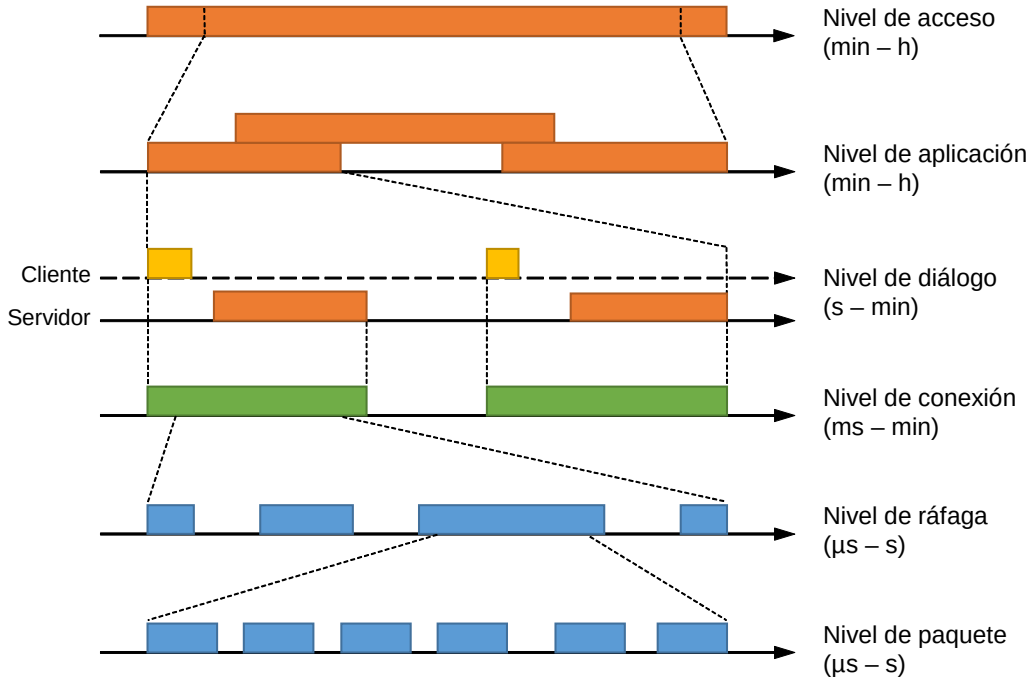

Figura 2.3: Niveles de actividad en comportamientos de usuario y aplicación

2.2.3. Modelos teóricos

Los modelos de tráfico pueden ser de tiempo continuo o discreto. En los modelos de tiempo continuo se basan en procesos estocásticos para representar tasa que varían en

el tiempo $X(t)$ o bien conjunto de tiempos de llegadas de paquetes $\{t_1, t_2, ...\}$. En los modelos de tiempo discreto se utilizan procesos estocásticos X_n para representar tasas de fuente muestreadas a lo largo de tiempos discretos $n = 1, 2, ...$ Desde el punto de vista conceptual, ambos tipos de modelos son muy similares, por lo que el uso de uno u otro depende del tipo de análisis que se quiera realizar y de la complejidad asociada a cada tipo de modelo de tráfico.

En esta tesis doctoral se propone el uso de modelos de tráfico basado en modelos de fuentes de tipo ON/OFF, cuyos modelos teóricos más relevantes en la literatura se describen a continuación.

2.2.3.1. Modelo simple de fuentes ON/OFF

El uso de los modelos basados en fuentes de tipo ON/OFF han gozado de gran popularidad debido a que reproducen algunas propiedades de tráfico en ráfagas de las redes de datos. El modelo de fuentes ON/OFF se compone de dos estados fundamentales: un periodo de actividad (ON), y un periodo de inactividad (OFF) donde la fuente no transmite datos. Un claro ejemplo de aplicación de este modelo se encuentra en una conversación, donde los participantes o bien se encuentran hablando o escuchando. En la figura 2.4 se muestra como el estado de una única fuente ON/OFF puede ser representado por una cadena de Markov con dos estados.

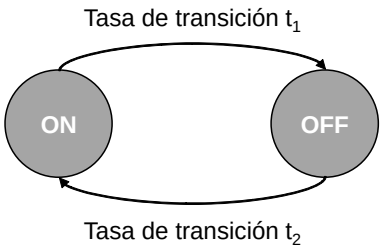

Figura 2.4: Modelo ON/OFF simple con tasas de transición t_1 y t_2

El diseño y desarrollo de estos modelos depende de la descripción y características de las entidades de tráfico involucradas desde el nivel de enlace hasta el de aplicación. Este modelo tienda a ser utilizado cuando se requiere tener en cuenta el comportamiento de escalado de tráfico de red. Por ejemplo, el análisis de tráfico de red a nivel IP se realiza de forma predominante mediante modelos de fuente ON/OFF.

En la mayoría de los casos, el tráfico no se modela mediante una única fuente de tipo ON/OFF, sino mediante la superposición de un conjunto de fuentes [Baiocchi et al., 1991]. Un modelo de *flujo fluido* basado en múltiples fuentes ON/OFF compartiendo un único buffer fue uno de los primero modelos propuestos para la caracterización de flujos de paquetes de red [Anick et al., 1982]. Posteriormente, en [Jain and Routhier, 1986] se

propone un modelo de este tipo de fuentes para modelar tráfico de redes Local Area Network (LAN). En [Willinger et al., 1997] se extiende el modelo y se muestra que bajo ciertas condiciones la superposición de este tipo de fuentes puede caracterizar tráfico de naturaleza autosimilar. En [Taqqu et al., 1997] se demuestra cómo se puede modelar tráfico autosimilar mediante la superposición de muchas fuentes de tipo ON/OFF con una alta variabilidad. Por esta razón, se pueden caracterizar el tráfico de una red de datos mediante el uso de muchas fuentes ON/OFF.

2.2.3.2. Modelos de Poisson modulados por Markov

Los modelos de Markov modulados constituyen una generalización de los modelos previos, ya que, como se ha mencionado anteriormente, los modelos de fuentes ON/OFF pueden ser expresados como un caso particular de un proceso de Markov modulado con únicamente dos estados.

Los modelos de tráfico basados en procesos de Markov modulados también son ampliamente utilizados para redes de conmutación de paquetes, ya que permite caracterizar tráfico *a ráfagas*. Un Markov Modulated Process (Proceso de Markov Modulado) (MMP) utiliza un proceso de Markov auxiliar, en el cual, en función del estado actual del proceso de Markov se controla la distribución de probabilidad de tráfico. En otras palabras, un MMP es un proceso doblemente estocástico que utiliza una cadena de Markov para definir la distribución de probabilidad de cada estado [Schwartz, 1996].

Un Markov Modulated Process (Proceso de Poisson modulado por Markov) (MMPP) es un caso particular de un MMP que usa procesos de Poisson modulados por una cadena de Markov, es decir, el ritmo de generación de llegadas cambia según el estado en el que nos encontremos en la cadena de Markov (a un estado s_k le corresponde un proceso de Poisson con una media λ_k). La figura 2.5 se presenta un ejemplo de un proceso de llegadas (por ejemplo de celdas a una red ATM) que utiliza un modelo de Markov modulado. Como se observa en la figura, las llegadas siempre se generan siguiendo un proceso de Poisson, pero con una tasa λ_i que depende del estado i en el que se encuentre la cadena de Markov (caja inferior).

Una de las grandes ventajas de este tipo de modelos es que pueden ser fácilmente representados mediante matrices, lo cual se traduce en una relativa simplicidad que habilita el uso de métodos analíticos para su evaluación. No obstante, la complejidad computacional de este tipo de modelos aumenta considerablemente en escenarios donde se consideran gran número de fuentes [Shroff and Schwartz, 1998].

2.2.3.3. Modelos de fuentes ON/OFF *heavy-tail*

Los procesos simples de fuentes ON/OFF descritos en la sección 2.2.3.1 suelen relacionarse con distribuciones de varianza finita para los tiempos de actividad e inactividad.

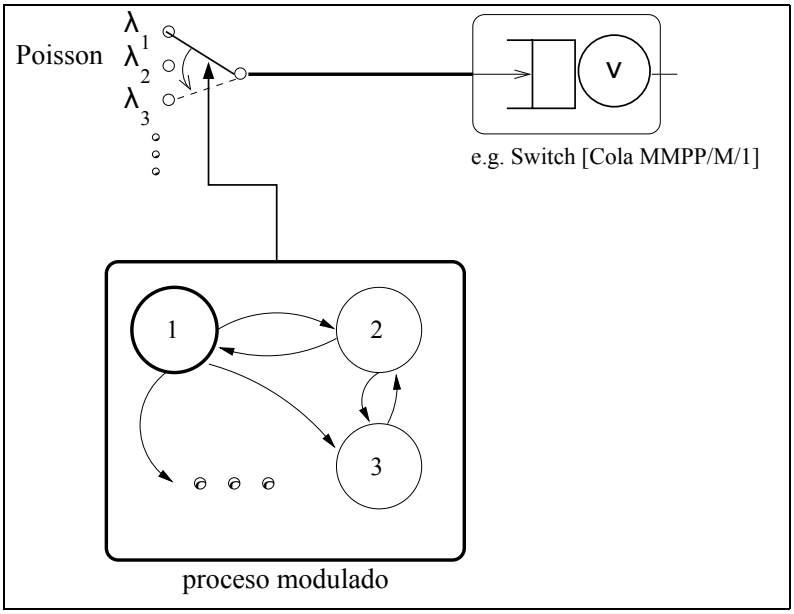

Figura 2.5: Proceso de llegadas MMPP en el contexto de un modelo de colas

Este hecho provoca que cuando se agregan múltiples fuentes, el tráfico caracterizado no presente algunas propiedades de Long Range Dependency (Dependencia a Largo Plazo) (LRD) propias de las redes de datos. No obstante, cuando las distribuciones de los periodos de actividad e inactividad de las fuentes siguen distribuciones *heavy-tail*, entonces los procesos pueden considerarse LRD [Heath et al., 1998]. A su vez, la agregación de múltiples fuentes ON/OFF de *heavy-tail* también muestra una LRD [Likhanov et al., 1995, Willinger et al., 1997, Willinger et al., 1998]. Estas características se corresponden con las aplicaciones reales de Internet, ya que éstas siguen distribuciones *heavy-tail* y la agregación de su tráfico exhibe un comportamiento de LRD [Paxson and Floyd, 1995]. En [GGP98] se demuestra la validez de modelos basados en fuentes ON/OFF *heavy-tail* para caracterizar tráfico LRD.

En [Schwefel and Lipsky, 2001, Schwefel and Lipsky, 1999] se nombra el modelo *N-Burst*, el cual considera el total de tráfico generado por un conjunto de N fuentes ON/OFF que siguen una distribución de *heavy-tail* para los periodos de actividad. Modelos similares se encuentran descritos en [Greiner et al., 1999, Hatem et al., 1997, Robert and Le Boudec, 1997].

2.2.3.4. Otros modelos

Anteriormente se han descrito modelos de tráfico que caracterizan el tráfico de red desde el punto de vista de las fuentes, es decir modelando el tráfico generado por los *hosts* conectados a la red. No obstante existen infinidad de modelos que pueden ser utilizados

para aproximar algunos modelos más específicos, como es el caso de la superposición de múltiples fuentes de tipo ON-OFF ($N \to \infty$) bajo ciertas circunstancias.

En [Likhanov et al., 1995] se analiza el tráfico generado por múltiples fuentes de tipo ON/OFF con distribuciones *heavy-tail* para los periodos de actividad o inactividad. Los autores demuestran como a medida que el número de fuentes aumenta, el proceso resultante se acerca a ocupación de un servidor de una cola M/G/∞ en los que los tiempos de servicio también siguen una distribución *heavy-tail*. Además, si la distribución de tiempo de servicio de una cola M/G/∞ es de tipo *heavy-tail*, el proceso que define el número de servidores ocupados es de tipo LRD.

Los modelos de tipo Fractional Brownian Motion (Movimiento Browniano Fracccional) (FBM) se basan en procesos estrictamente auto-similares de tipo gaussiano y que cumplen ciertas condiciones. Se demuestra que, para el caso discreto, la autocorrelación normalizada de la secuencia de incrementos, también denominada ruido gaussiano fraccional, presenta LRD. En [Brichet et al., 1996] se demuestra que si se superponen N fuentes ON/OFF de tipo *flujo fluido* con distribuciones *heavy-tail*, el proceso resultante converge a un proceso gaussiano a medida que el número de fuentes crece.

De la misma forma, existen otros modelos que pueden ser utilizados para la aproximación de la superposición de fuentes de tráfico. Por ejemplo, en [Mikosch et al., 2002] se muestra como existen algunas circunstancias en las que esa agregación de fuentes de distribución *heavy-tail* no converge a un FBM. En el caso en el que las distribuciones del tamaño de las ráfagas sea comparativamente mucho más *heavy-tail* que la tasa de llegadas de ráfagas, existe una convergencia a distribuciones denominadas *α-stable Lévy Motion*.

Existen otros modelos en la literatura que gozan de gran popularidad (modelo AutoRegressive Integrated Moving Average (Media Móvil Integrada AutoRegresiva) (ARIMA), análisis multifractal, etc.), que también se utilizan para aproximar modelos de tráfico [Park and Willinger, 2000].

2.2.4. Mezcla de tráfico de aplicaciones

El tráfico en redes de comunicaciones de datos se genera por un conjunto de diferentes aplicaciones y/o servicios. Es por ello que, un desafío importante a la hora de caracterizar y modelar el tráfico de Internet, se encuentra en la identificación de las aplicaciones que predominan en la red (por ejemplo, navegación web, video online, etc.), así como, de la cantidad de tráfico contribuido por las mismas. Además, la composición de la mezcla de tráfico generado va evolucionando a medida que aparecen y desaparecen nuevas aplicaciones y servicios. Por esta razón, es muy importante que la caracterización y modelado del tráfico de Internet sea flexible a la hora de modificar el conjunto de aplicaciones considerado, así como sus propias características intrínsecas.

En la literatura existen numerosos estudios que investigan esta composición de tráfico de Internet, con el objetivo de identificar aquellas aplicaciones o servicios que más tráfico generan [Labovitz et al., 2011, Kihl et al., 2010, Vu-Brugier, 2009, Maier et al., 2009]. En [Labovitz et al., 2011] se realiza un estudio a gran escala del tráfico inter-dominio del que se extrae la conclusión que el tráfico predominante es el tráfico web, con una aportación de algo más del 50 %. No obstante, tras realizar una inspección de paquetes en profundidad, se descubre que gran parte de este tráfico se debe a servicios de video sobre Internet. Utilizando esta misma técnica, también se extrae que la aportación de aplicaciones de compartición de ficheros (por ejemplo, aplicaciones P2P como BitTorrent) son responsables de la generación de cuotas de tráfico cercanas al 20 % [Maier et al., 2009]. Estos datos son consistentes con un estudio reciente [Feknous et al., 2014], donde se analiza el tráfico capturado en operadores europeos en redes fijas (Digital Subscriber Line (Línea de Abonado Digital) (DSL) y FTTH) y móviles. En [Cisco, 2014] también se presenta un detallado análisis de las aportaciones de cada tipo de aplicación al tráfico de Internet consumido en España, donde no sólo se identifican las aplicaciones más representativas, sino además se estima la cantidad de tráfico consumido por estas aplicaciones a lo largo del mes.

En general, en la gran mayoría de referencias bibliográficas recientes se identifican el mismo conjunto de aplicaciones responsables de la generación de la mayor parte de tráfico de Internet [Cisco, 2014, Feknous et al., 2014, Katsaros et al., 2012, Wamser et al., 2011]:

- **Web (y otros):** esta categoría está compuesta principalmente por tráfico web. No obstante, también incluye las pequeñas aportaciones del tráfico de otro tipo de aplicaciones como son el correo electrónico, la mensajería instantánea y otros tipos de datos que no se encuentran incluidas en otras categorías.

- **Compartición de ficheros:** incluye tráfico de tipo de aplicaciones Peer-to-Peer (P2P) de todos los sistemas conocidos de este tipo, como por ejemplo *BitTorrent* y *eDonkey*, asi como el tráfico de los sistemas de compartición de ficheros basados en web.

- **Video sobre Internet:** esta categoría incluye multitud de tipos de videos sobre Internet, incluyendo videos cortos (por ejemplo, *YouTube*), videos de formato largo (por ejemplo, *Hulu* o *Netf*), video en directo, video de Internet al televisor, visionado de webcams, etc.

- **Juegos en red:** esta categoría incluye cualquier tipo de juegos en Internet, desde el juego casual online (por ejemplo, a través de *Facebook*) y juegos online, ya sean de consolas u ordenador.

En la siguiente sección, se presenta un análisis de los modelos de tráfico de diferentes aplicaciones de Internet disponibles en la literatura, con objeto de buscar un modelo de tráfico global compuesto por modelos de tráfico para cada tipo de aplicación representativa de Internet.

2.2.5. Modelos de navegación web

El tráfico web incluye todo el tráfico generado por el protocolo Hypertext Transfer Protocol (Protocolo de Transferencia de Hipertexto) (HTTP) durante una sesión de navegación web (por ejemplo, a través de aplicaciones como *Firefox*, *Chrome*, etc.). Una página web consiste en un objeto principal y un conjunto de elementos de la página que se encuentran en un fichero adicional (por ejemplo, imágenes, *scripts*, etc.). En función de las diferentes versiones existentes de HTTP [Berners-Lee et al., 1996, Fielding et al., 1999] se producen más o menos conexiones TCP, lo cual tiene un impacto en el tipo de tráfico generado.

En la literatura, los modelos suelen basarse en las trazas obtenidas a partir del servidor, del cliente o del análisis de trazas de red, siendo esta última la más popular en los últimos años. En general, se realizan medidas a partir de las trazas a nivel de paquete de una red que transporte tráfico web, como por ejemplo, una LAN. Los modelos de tráfico pueden diferir en el nivel de actividad considerado (sección 2.2.2), de forma que pueden ser clasificados de la siguiente manera:

- En [Vicari, 1997, Färber et al., 1999] se analiza el comportamiento de los acceso (*dial-in*) de los usuarios en redes LAN.
- En [Vicari, 1997, Färber et al., 1999, Reyes-Lecuona et al., 1999] se realiza un análisis en el que hacen una distinción entre sesiones y *subsesiones*, proponiendo distribuciones tanto para el tiempo entre llegadas, así como para la duración o tamaño de la sesión.
- En [Deng, 1996, Mah, 1997, Choi and Limb, 1999] no se realiza la distinción anterior de identificar sesiones, sino que se supone un nivel de actividad de aplicación infinito.

En [Vicari, 1997] se introduce el término de *subsesión*, la cual se define como el periodo de tiempo en el que un usuario genera tráfico web sin que exceda de un umbral de tiempo predeterminado *time-out*. Este umbral de tiempo se selecciona para detectar cuando un usuario puede estar leyendo una web sin que haya pedido una página nueva, siendo el valor seleccionado por el autor de 3 segundos. En la mayoría de los casos se podría asumir que las sesiones y subsesiones son idénticas. En los resultados del estudio, una subsesión cuenta con 19,6 páginas web que contienen a su vez 4 elementos

en media. En este trabajo se modela el tamaño de una web completa y el tiempo entre dos peticiones web con distribuciones normalizadas de Pareto. No obstante, también se detallan algunas limitaciones de la elección de este modelo, pues se sobrerrepresenta la Short Range Dependency (Dependencia a Corto Plazo) (SRD) y no se caracteriza debidamente la naturaleza de LRD de este tipo de tráfico.

En [Färber et al., 1999] se analiza cómo escalan diferentes procesos estocásticos en función la agregación de tráfico producida por múltiples usuarios. Los autores comparan diferentes distribuciones durante la hora cargada frente a las trazas capturadas para el estudio, concluyendo que modelos basados en distribuciones de tipo hiperexponencial aportan buenas estimaciones.

En [Reyes-Lecuona et al., 1999] se presenta un modelo de tráfico más detallado que consiste en tres diferentes niveles de actividad: nivel de sesión, nivel de página y nivel de paquete. El nivel de sesión, definido por el tiempo entre llegadas de sesiones, se modela mediante un proceso de Poisson. El número de páginas por sesión lo caracterizan con una distribución de tipo lognormal. El modelo a nivel de página se define por el tiempo entre páginas y el tamaño de las mismas, y que se aproximan por distribuciones de tipo gamma y Pareto respectivamente. A nivel de paquete el tráfico se caracteriza por el tamaño de paquete con una distribución multimodal y tiempo entre llegadas de paquetes con una distribución exponencial.

Los estudios descritos [Deng, 1996, Mah, 1997, Choi and Limb, 1999] presentan un modelo basado en fuentes ON/OFF y que no tienen en cuenta los tiempos de inicio de sesión de los usuarios, sino el tiempo entre descargas de páginas web. Por lo tanto, el periodo ON contempla el tiempo necesario para la descarga de una página y el periodo OFF se corresponde con el tiempo que existe entre descargas de páginas. Este tiempo en el que la fuente se encuentra apagada también se le denomina *thinking time* debido a que el usuario se encuentra visualizando los contenidos descargados. Los resultados extraídos en [Deng, 1996] sugieren el uso de una distribución de tipo Weibull para los periodos ON y Pareto para los periodos OFF. Una limitación de este estudio reside en que no se tiene en cuenta la estructura de la página web solicitada.

En [Mah, 1997] se presenta un modelo mucho más detallado que el anterior que tiene en cuenta los diferentes elementos presentes en una página web. Además, el periodo de inactividad OFF también incluye los tiempos entre dos sesiones consecutivas de usuario. El autor propone algunas distribuciones para los periodos de actividad, así como para el tamaño del objeto principal de la página. No obstante, este estudio no tiene como objetivo proponer distribuciones para modelar los diferentes estados de la fuente de tráfico.

Posteriormente, en [Choi and Limb, 1999] se presenta un nuevo modelo de tráfico web basado en nuevos parámetros, que anteriormente no habían sido tenidos en cuenta

y que ha supuesto el punto de partida para muchos otros estudios. En la figura 2.6 se ilustra un modelo basado en fuentes ON/OFF que se ajusta al propuesto por Choi y Limb, ya que presenta las diferentes abstracciones de niveles de actividad para el caso de tráfico web. Estos autores modelan el periodo de actividad en función de los siguientes parámetros: número de objetos en línea, tiempo de llegada entre objetos en línea, tamaño del objeto principal, tamaño de los objetos en línea y tiempo de parada. Se modela cada uno de los parámetros que conforman el periodo ON mediante diferentes distribuciones de tipo lognormal, gamma y geométricas. El periodo de visionado (OFF) se modela mediante una distribución Weibull. A pesar de aportar los parámetros de cada una de las distribuciones correspondientes al periodo ON, los autores validan el modelo comparándolo con otros estudios de la literatura y concluyendo que, el periodo ON total se puede ajustar mediante una distribución de tipo Weibull.

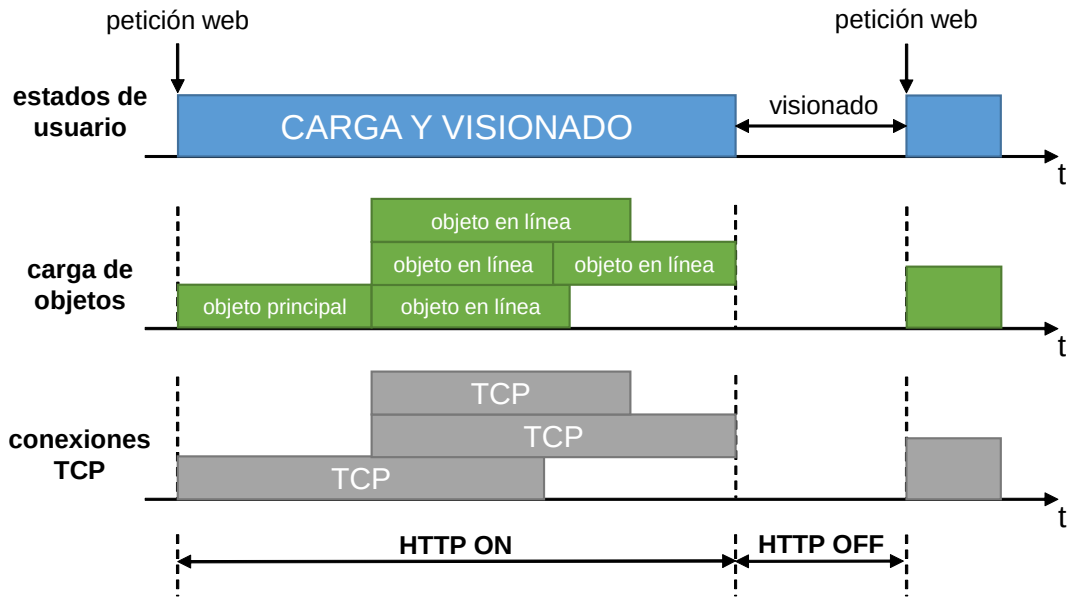

Figura 2.6: Modelo de tráfico WEB basado en comportamiento de usuario

En [Zhu et al., 2003] se proponen dos modelos de tráfico diferentes en función del tipo de red, móviles o fija. Los modelos propuestos en este trabajo tienen en cuenta diferentes niveles de actividad. Para el nivel de sesión, los autores definen unos tiempos de actividad e inactividad basados en distribuciones Weibull y exponencial respectivamente. Bajando de nivel, este estudio descompone el periodo de actividad en el número de páginas por sesión, tiempo de llegadas entre páginas y tamaño de páginas, modelados con distribuciones Weibull. Por último, en el nivel de objeto, modelan el tiempo entre llegadas de objetos mediante una distribución Weibull y el tamaño de objeto con una

distribución exponencial.

Lee y Gupta proponen un nuevo modelo de tráfico en [Lee and Gupta, 2007] que pretende abordar la limitación de modelos anteriores, donde no se tiene en cuenta que durante el periodo de actividad se puedan hacer múltiples peticiones de páginas web. El modelo propuesto también se encuentra basado en fuente de tipo ON/OFF, pero considerando el caso de que existan múltiples páginas dentro de una misma petición web. El periodo ON tiene en cuenta multitud de parámetros, entre los que destacan el tamaño de objeto principal (lognormal), el número (gamma) y tamaño de objetos en línea (lognormal), los tiempos de respuesta entre objetos en línea (Weibull). Para el periodo de visionado se utiliza una distribución de tipo lognormal.

En [Pries et al., 2012] se pone de manifiesto que el tráfico web de Internet cambia constantemente con el paso del tiempo, de forma que es necesario un ajuste en los modelos de tráfico en la literatura para que éstos puedan ser válidos en simulaciones y análisis. En este trabajo se parte de los datos recopilados del millón de páginas web más populares en la red. A partir de esta información utilizan los modelos de tráfico descritos en la literatura para escoger aquellas distribuciones más adecuadas y ajustar sus parámetros para que sean válidos con los datos reales de Internet. Siguiendo el modelo de Lee y Gupta [Lee and Gupta, 2007], se tiene en cuenta la posibilidad de que puedan existir varias peticiones de páginas dentro de un mismo periodo de actividad ON, es decir se define una distribución de tipo lognormal para el número de objetos principales. De la misma forma, también se tienen en cuenta otros parámetros para el periodo ON, como por ejemplo, el tamaño de objeto principal (Weibull), el número de objetos en línea (exponencial) y su longitud (lognormal). El periodo de inactividad o de visionado se ajusta a una distribución de tipo lognormal.

2.2.5.1. Resumen

A partir del análisis de los modelos de tráfico web de la literatura se puede concluir que cada uno de ellos ha tratado de continuar con los descubrimientos de los anteriores e intentando solucionar algunas de las limitaciones que tenían presentes. Por esta razón se puede decir que no son modelos disruptivos entre sí, sino que gozan de una naturaleza continuista. En la tabla 2.8 se muestra una visión general de los modelos de tráfico web más recientes para aquellos parámetros que pueden ser comparados.

En esta tesis doctoral se propone el uso del modelo propuesto en [Pries et al., 2012] ya que aborda algunas de las limitaciones presentes en modelos de tráfico previos. Tiene en cuenta la posibilidad de la existencia de múltiples objetos principales (no se encuentra descrito en la tabla 2.8). Además, este modelo ha sido desarrollado y ajustado a partir de los sitios web más visitados, y no a partir de las trazas de algún escenario limitado, como por ejemplo, las trazas correspondientes de un campus universitario o de pequeños

Referencia	Tamaño de objeto principal	Tamaño de objeto en línea	Número de objetos en línea	Tiempo de visionado
[Mah, 1997]	Pareto $\alpha = 0,85 - 0,97$ $med. = 2 - 2,4$ kB	Pareto $\alpha = 1,12 - 1,39$ $med. = 1,2 - 2$ kB	- $E(X) = 2,8 - 3,2$ $med. = 1$	- $E(X) = 1000 - 1900$ s $med. = 15$ s
[Choi and Limb, 1999]	Lognormal $E(X) = 10$ kB $med. = 6$ kB $\sigma = 25$ kB	Lognormal $E(X) = 7,7$ kB $med. = 2$ kB $\sigma = 126$ kB	Gamma $E(X) = 5,5$ $med. = 2$ $\sigma = 11,4$	Weibull $E(X) = 39,5$ s $med. = 11$ s $\sigma = 92,6$ s
[Lee and Gupta, 2007]	Lognormal $E(X) = 11,9$ kB $\sigma = 38$ kB	Lognormal $E(X) = 12,5$ kB $\sigma = 116$ kB	Gamma $E(X) = 5,07$	Lognormal $E(X) = 39,7$ s $\sigma = 324,92$ s
[Pries et al., 2012]	Weibull $E(X) = 31,6$ kB $med. = 19,5$ kB $\sigma = 49,2$ kB	Lognormal $E(X) = 23,9$ kB $med. = 10,3$ kB $\sigma = 128$ kB	Exponential $E(X) = 31,92$ $med. = 22$ $\sigma = 237,65$	Lognormal $E(X) = 39,7$ s $\sigma = 324,92$ s

Tabla 2.8: Visión general de los modelos de tráfico web

Parámetro del modelo	**Distribución**	**Parámetro 1**	**Parámetro 2**
Número de objetos principales	Lognormal	$\mu = 0,473844$	$\sigma = 0,688471$
Tamaño de objeto principal	Weibull	$\alpha = 28242,8$	$\beta = 0,814944$
Número de objetos en línea	Exponencial	$\mu = 31,9291$	-
Tamaño de objeto en línea	Lognormal	$\mu = 9,17979$	$\sigma = 1,24646$
Tiempo de visionado	Lognormal	$\mu = -0,495204$	$\sigma = 2,7731$

Tabla 2.9: Parámetros del modelo de tráfico web de [Pries et al., 2012]

Internet Service Providers (Proveedores de Servicios de Internet) (ISPs). En definitiva, este modelo no sólo tiene en cuenta la popularidad de las páginas en Internet más visitadas, sino que además no se encuentra limitado a los usuarios de una zona geográfica específica.

En la tabla 2.9 se muestra los parámetros del modelo de tráfico web seleccionado para esta tesis doctoral [Pries et al., 2012]. Mencionar que el tiempo de visionado del modelo se encuentra basado en el ajuste realizado en [Lee and Gupta, 2007].

2.2.6. Modelos de compartición de ficheros

Los servicios de compartición de ficheros suelen encontrarse basadas en aplicaciones basadas en File Transfer Protocol (Protocolo de Transferencia de Archivos) (FTP), P2P. No obstante, la descarga de contenidos digitales suele realizarse mediante sistemas

basados en tecnologías P2P o bien mediante enlaces de descarga directa (HTTP).

Existen numerosas plataformas o aplicaciones de compartición de ficheros, cuya popularidad depende de diversos factores como por ejemplo el tipo de contenido o incluso, la dureza de las leyes de derechos intelectuales de los países donde se encuentran los usuarios. No obstante, es cierto que las aplicaciones más relevantes suelen compartir ciertas características en cuanto al tipo de contenido disponible, su tamaño y popularidad [Cunche et al., 2012].

A continuación se presentan modelos de tráfico más relevantes para las aplicaciones de compartición de ficheros. En primer lugar, se presentan los modelos basados en FTP ya que son más sencillos desde el punto de vista de la arquitectura e interacciones entre entidades. Posteriormente, se analizan algunos de los modelos más relevantes de aplicaciones P2P. Por último se presenta la distribución de contenidos a través de servicios de alojamiento de archivos.

2.2.6.1. Modelos FTP

El tráfico de transferencia de ficheros mediante FTP [Postel and Reynolds, 1985] se caracteriza por sesiones en las que los usuarios realizan una secuencia de transferencias de contenidos digitales, separadas por un tiempo determinado (Figura 2.7). Este tiempo, a menudo denominado *tiempo de visionado* en referencia a los modelos web, se define como el tiempo entre el final de una transferencia de un fichero y la petición de transferencia del siguiente.

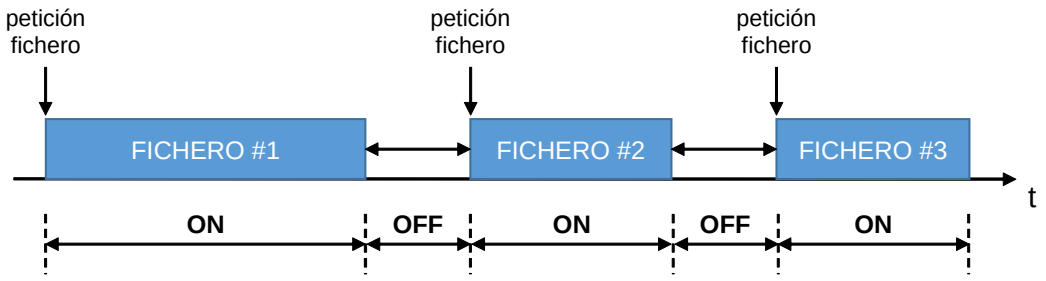

Figura 2.7: Modelo conceptual de tráfico FTP

En [Srinivasan et al., 2008] se muestra una metodología de evaluación de tecnologías de red basada en un conjunto de modelos de tráfico, entre los que se encuentra un modelo de FTP. Este modelo sigue el enfoque anterior, ya que se caracteriza por periodos de actividad definido por el tamaño del fichero a descargar, y periodos de inactividad que se corresponden con el tiempo entre descargas. El tamaño de fichero se ajusta mediante una distribución lognormal truncada con una media de 2 MB ($\sigma = 0{,}35$, $\mu = 14{,}45$,

$max = 5MB$). El tiempo de inactividad sigue una distribución exponencial con media de 180 segundos ($\lambda = 0{,}006$).

En [Luo and Marin, 2005] se modela tráfico realista de Internet mediante un conjunto de aplicaciones, entre las que se encuentra FTP. De forma análoga al modelo anterior, se consideran periodos de actividad e inactividad. Además, se modelan el número de conexiones de datos FTP que existen por sesión mediante una distribución de Pareto ($\alpha = 1{,}0595, k = 3$). El tamaño de los datos transferidos en los periodos de actividad se ajusta mediante la combinación de una distribución de Pareto ($p = 0{,}15, \alpha = 1{,}15, k = 10000$) y exponencial ($p = 0{,}85, tasa = 0{,}00052$). El tiempo entre llegadas de los periodos de transferencia de datos se caracterizan mediante distribuciones gamma en función de si el cliente FTP realiza las transferencias de forma manual ($k = 0{,}227, \theta = 73{,}962$) o automática ($k = 202{,}04, \theta = 0{,}079$).

En la tesis doctoral [Svoboda, 2008] se modela el tráfico FTP mediante la caracterización del tamaño de fichero, las llamadas de servicio por usuario y el número de usuarios. El tamaño de los ficheros se caracteriza mediante una distribución lognormal ($\mu = 8{,}5534, \sigma = 2{,}112$). Según el autor, la distribución de llamadas de servicio de usuario se ajusta a una distribución exponencial ($mu = 6{,}542$) y es independiente del tipo de tecnología de acceso.

A pesar de que en su momento las aplicaciones FTP contaron con cierta popularidad, en la actualidad su uso ha disminuido drásticamente en detrimento de otro tipo de aplicaciones de compartición de ficheros (por ejemplo, P2P). Por esta razón existen muchas otras referencias bastante antiguas en la literatura acerca del tráfico FTP, entre las que destacan algunos de los primeros modelos de tráfico descritos en [Paxson, 1994, Cáceres et al., 1991]. Debido a la poca popularidad con la que cuentan este tipo de aplicaciones en la actualidad, este tipo de modelos de tráfico no son representativos para caracterizar a este tipo de aplicaciones.

2.2.6.2. Modelos P2P

Una red P2P es una red de comunicaciones de ordenadores que permite el intercambio directo de información, en cualquier formato, entre los usuarios interconectados. Algunos ejemplos de aplicaciones de compartición de ficheros son: *Napster*, *Kazaa*, *Gnutella* o *BitTorrent*.

Las aplicaciones P2P pueden tener arquitecturas centralizadas, descentralizadas o híbridas, lo cual tiene un impacto directo en cómo es el tráfico que se genera entre *peers*. Tradicionalmente, las descargas se realizan entre *peers*, la señalización puede seguir un esquema cliente-servidor (por ejemplo, *Napster*).

Tamaño de ficheros. Las redes P2P son heterogéneas en cuanto a los diferentes tipos de contenidos y el tamaño de los ficheros existentes. En [Aidouni et al., 2009] se presenta un análisis básico de los contenidos presentes en un servidor P2P *eDonkey*. Por un lado, los autores presentan la distribución de los tamaños de los contenidos presentes en el servidor, tal y como se muestra en la figura 2.8. En la figura se observa la existencia de gran cantidad de pequeños ficheros, probablemente ficheros de música e imágenes. Además, existen claros picos en los tamaños de 700 MB, correspondiente al tamaño típico de un CD-ROM, y en algunas porciones de este valor, como es el caso de 350 MB, 230 MB, y 175 MB. Además, también existe un pico en el valor de 1.4 GB, correspondiente al tamaño de un DVD. El pico situado en 1 GB puede deberse a que algunos usuarios dividen ficheros de gran tamaño en ficheros menos de 1 GB de tamaño. Estos resultados son coherentes con estudios anteriores. En el trabajo descrito [Leibowitz et al., 2002] se describe una distribución de archivos similar a la anterior, donde también predominan pequeños archivos de audio. Además, los autores también comentan la gran popularidad con la que cuentan los ficheros correspondientes a películas.

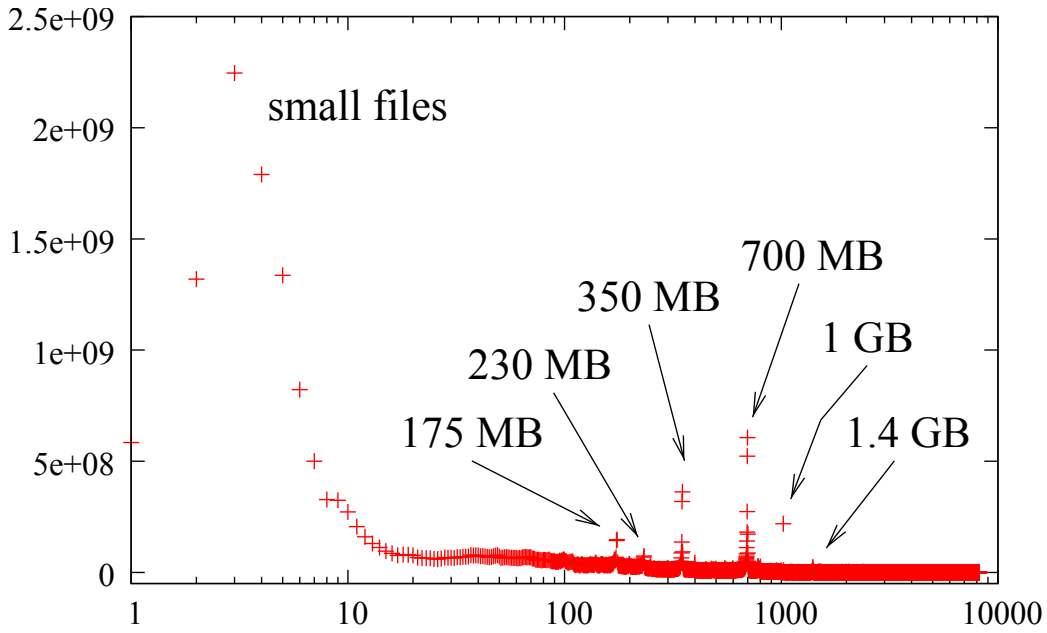

Figura 2.8: Distribución de tamaño de ficheros en red P2P [Aidouni et al., 2009]

Popularidad de contenido. Otro parámetro a tener en cuenta a la hora de caracterizar los contenidos en una red P2P es la popularidad de los mismos. Los tamaños de los contenidos más populares no tiene porqué corresponderse con la distribución de tamaños de ficheros disponibles en la red. En [Aidouni et al., 2009] se estudia la popularidad como el número de nodos que piden o comparten los contenidos de una red

P2P. En este caso de estudio, la gran mayoría de ficheros (más de 5 millones) cuentan una popularidad menor o igual a 10, mientras que unos pocos ficheros (casi 50) tienen una popularidad mayor de 50000.

En [Hawa et al., 2012] se realiza un estudio de la red P2P *BitTorrent* durante el año 2008 y 2011. En primer lugar, este estudio realiza una clasificación de los contenidos disponibles en la red en diferentes tipos: audio, video, archivos, imágenes de CD y documentos. Para cada uno de los tipos de contenidos se analizan el número relativo de ficheros existentes en el sistema: audio (34 %), video (11,2 %), archivos (24,4 %) imágenes de CD (0,3 %) y documentos (29,8 %). No obstante, el volumen relativo compartido de los diferentes tipos de contenidos es completamente diferente: audio (6,8 %), video (67,4 %), archivos (19,3 %) imágenes de CD (6 %) y documentos (0,7 %). En estas cifras se puede comprobar cómo la frecuencia del tipo de contenido disponible en la red difiere de la popularidad del mismo.

Aplicación de ejemplo: *BitTorrent*. De entre todas las aplicaciones P2P existentes, la más popular en la actualidad es *BitTorrent*. *BitTorrent* no sólo hace referencia a los sistemas de compartición de ficheros, sino que además es un protocolo para la distribución de contenidos diseñado para replicar de forma rápida y efectiva los datos entre sus clientes, denominados *peers*. A diferencia con otros protocolos de este tipo, *BitTorrent* no proporciona ningún mecanismo de búsqueda de contenidos. Un cliente que se encuentre interesado en la descarga de un contenido, ha de descargase en primera instancia un archivo de metadatos, denominado *torrent*, el cual tiene la información para que el *peer* pueda iniciar una comunicación con el resto de *peers* que disponen del contenido deseado. Esta red de clientes conectados entre sí para descargar o subir un contenido se le denomina enjambre (*swarm*). En un *BitTorrent swarm* existen tres entidades bien diferenciadas: *trackers*, *seeders* y *leechers*. Los primeros son una entidad central que guarda una lista de los *peers* conectados y estadísticas sobre la evolución del enjambre. Los *peers* activos pueden ser categorizados en *seeders* o *leechers*. Los *leechers* son nodos activos que se encuentran en la fase de descarga del contenido. Los *seeders* ya se han descargado el contenido por completo pero continúan compartiéndolo de forma altruista con otros nodos.

Modelos de tráfico. Debido a que las aplicaciones P2P son más recientes que otras aplicaciones tradicionales de Internet, existen escasas referencias en la literatura que modelen de forma estocástica el tráfico. La gran mayoría de ellas analizan estadísticamente el tráfico en diferentes escenarios.

En [Erman et al., 2005] y [Erman et al., 2006] se presenta un modelo y evaluación de las características de sesión de tráfico capturado de la aplicación P2P, *BitTorrent*. Los principales resultados del estudio es que los tiempos entre llegadas de sesiones pueden

Aplicación	Métrica	x_{corte}	Prob.	Body	Tail
BitTorrent	Volumen	-	-	Weibull$(0,084; 0,027)$	-
	Duración	8,83	0,213	Weibull$(0,184; 1,408)$	Weibull$(0,00015; 1,39)$
	IAT	-	-	Weibull$(0,709; 0,443)$	-
Napster	Volumen	-	-	Pareto$(25; 0,702)$	-
	Duración	-	-	Weibull$(0,641; 0,363)$	-
	IAT	2s	0,8	Weibull$(2; 237; 0,775)$	Pareto$(2; 2,188)$
eDonkey	Volumen	$10^{5,249}$	0,9	Weibull$(0,019; 0,556)$	Exponencial$(1,63 \cdot 10^6)$
	Duración	-	-	Weibull$(0,05; 0,706)$	-
	IAT	-	-	Weibull$(0,644; 0,874)$	-
Gnutella	Volumen	$10^{5,003}$	0,95	Pareto$(10^{2,4}; 1,11)$	Weibull$(0,0002; 2,703)$
	Duración	10s	0,5	Weibull$(0,234; 1,236)$	Weibull$(0,0018; 1,270)$
	IAT	-	-	Weibull$(0,699; 0,493)$	-
Fasttrack	Volumen	$10^{3,39}$	0,5	Weibull$(0,0007; 1,152)$	Pareto$(10^{3,39}; 0,52)$
	Duración	8,85s	0,54	Weibull$(0,281; 1,155)$	Pareto$(8,85; 0,72)$
	IAT	-	-	Weibull$(0,694; 0,477)$	-

Notación: Weibull(α, β), Pareto(k, α), Exponencial(μ).

Tabla 2.10: Modelos de tráfico para aplicaciones P2P de [He et al., 2007]

ser ajustados a funciones hiper-exponenciales, mientras que las duraciones y tamaños de las sesiones se pueden modelar mediante distribuciones lognormal. Los ajustes a funciones se realizan para 13 medidas de tráfico diferentes, dando por tanto a diferentes ajustes de parámetros en las distribuciones. Por esta razón, la principal conclusión de este estudio no es la aportación de un modelo de tráfico P2P, sino detallar los tipos de distribuciones que se ajustan mejor a las capturas realizadas.

En [He et al., 2007] se analizan y modelan algunas de las aplicaciones P2P más utilizadas a partir de la captura y análisis de tráfico de red en un enlace transoceánico. Las aplicaciones P2P son las siguientes: *Napster*, *BitTorrent*, *eDonkey*, *Gnutella* y *Fasttrack*. Los modelos caracterizan el tráfico de cada aplicación a nivel de flujo a partir de distribuciones de probabilidad para el volumen tráfico, las duraciones de las conexiones y los tiempos entre llegadas (InterArrival Time (Tiempo Entre Llegadas) (IAT)) de conexiones. En la tabla 2.10 se muestran las distribuciones utilizadas para cada aplicación y característica de tráfico.

En [Basher et al., 2008] se comparan dos de las aplicaciones más relevantes de la compartición de ficheros P2P: *BitTorrent* y *Gnutella*. En este estudio se consideran métricas a nivel de flujo para desarrollar un modelo de tráfico que pueda ser utilizado para sintetizar cargas de tráfico de aplicaciones P2P. En las ecuaciones 2.15 y 2.16 se describe el modelo para aplicaciones P2P mediante distribuciones para el tamaño de

flujo y para el tiempo entre llegadas de flujos.

$$F_{P2P}(S) = \begin{cases} 1 - e^{-(\frac{S}{1,36})^{0,81}} & \text{si } S < 4KB \\ 1 - (\frac{0,005}{S})^{0,35} & \text{si } 4KB \leq S \leq 10MB \\ 1 - (\frac{400}{S})^{1,42} & \text{si } S > 10MB \end{cases} \tag{2.15}$$

$$F_{P2P}(IAT) = \begin{cases} 1 - e^{-(\frac{IAT}{0,35})^{0,87}} & \text{si } IAT \leq 0,1sec \\ 1 - e^{-(\frac{IAT}{0,45})^{0,65}} & \text{si } 0,1sec \leq IAT \leq 1sec \\ 1 - (\frac{0,18}{IAT})^{0,97} & \text{si } IAT > 1sec \end{cases} \tag{2.16}$$

En [Glaropoulos et al., 2014] se define un modelo para el estudio del nivel de ocupación en redes inalámbricas 802.11 basado en modelos de tráfico de un conjunto representativo de aplicaciones: Domain Name System (Sistema de Nombres de Dominio) (DNS), Web, FTP, P2P, Voice over IP (Voz sobre IP) (VoIP) and video. Para la aplicación P2P utilizan un modelo descrito en [Çiflikli et al., 2010] para el caso particular de *BitTorrent*. Este modelo de tráfico se encuentra definido a nivel de paquete IP y utiliza una distribución exponencial ($\lambda = 512$) para el tamaño de paquetes y una distribución Weibull ($\alpha = 0,53; \beta = 0,13532$) para el tiempo entre llegadas de paquetes.

En la literatura también existen algunos modelos de tráfico P2P complejos que tienen en cuenta algunas características adicionales a las consideradas en modelos anteriores. Por ejemplo, en [Xu et al., 2014] se desarrolla un modelo de tráfico P2P, donde se consideran el número de *peers*, el factor de localización y distancia de la red. Este modelo genera matrices de tráfico P2P teniendo en cuenta las diferentes entidades existentes en la red y los diferentes estados en los que se puede encontrar un cliente. No obstante, hay que reseñar que los autores describen el modelo pero no aportan datos numéricos del mismo, por lo que no puede ser utilizado para generar tráfico de red de este tipo. Otro ejemplo se encuentra en [Lai et al., 2014], donde se analiza y caracteriza únicamente la fase inicial del proceso de compartición de ficheros en redes de tipo *BitTorrent*. En [Yang and De Veciana, 2004] se descompone la caracterización de la capacidad de un servicio P2P como *BitTorrent* en dos regímenes. El primer hace referencia a la fase transitoria en la que el sistema intenta hacer frente a ráfagas de demandas de contenidos. El segundo régimen es la fase estacionaria, donde el sistema escala en función de la demanda de los clientes. Los autores definen los modelos para ambas regímenes pero no ofrecen datos numéricos para poder generar tráfico P2P.

2.2.6.3. Servicios de alojamiento de archivos (*cyberlockers*)

En los últimos años han proliferado algunos servicios de alojamiento de archivos (e.g. *RapidShare*, *Megaupload*, etc.) que han sustituido parcialmente a los sistemas y

aplicaciones de compartición de ficheros [Gehlen et al., 2012]. Por ejemplo, en [Maier et al., 2009] ya se reporta que un único servicio de alojamiento consumía un 15 % del ancho de banda en una red residencial. Este servicio, también denominado descarga directa, mediante el cual un usuario puede subir, gestionar y compartir contenidos en un servidor remoto o nube. El caso de uso más común consiste en subir un contenido a un servidor de este servicio para posteriormente ser compartido con otros usuarios mediante un Uniform Resource Locator (Localizador de Recursos Uniforme) (URL) generada. El tráfico generado por este tipo de servicios es similar al tráfico web debido al protocolo subyacente (HTTP).

En [Mahanti et al., 2012] se realiza un amplio análisis de las características de tráfico de los principales servicios de alojamiento de ficheros en la red de un campus universitario. A nivel de flujo, los autores presentan un modelo para caracterizar los tamaños de los flujos de descarga de contenidos mediante distribuciones de tipo lognormal (ecuación 2.17). Este estudio ajusta tiempo entre llegadas de flujos según la ecuación 2.18. Además, también modela las duraciones y tasas de los flujos mediante combinaciones de distribuciones de tipo lognormal y gamma.

$$F_{CC}(S) = \begin{cases} LN(\mu = 6,02; \sigma^2 = 0,3) & \text{si } S < 329KB \\ LN(\mu = 10,2; \sigma^2 = 5,98) & \text{si } 329KB \leq S < 45MB \\ LN(\mu = 10,66; \sigma^2 = 1,07) & \text{si } 45MB \leq S \end{cases} \tag{2.17}$$

$$F_{CC}(\Delta) = \begin{cases} Gamma(a = 2,27; b = 5,62; c = 0,09) & \text{si } \Delta < 3sec \\ LN(\mu = 3,93; \sigma^2 = 2,15) & \text{si } 3sec \leq \Delta \end{cases} \tag{2.18}$$

A nivel de sesión, este estudio también presenta un modelo donde se caracteriza el volumen de transferencia de datos y el tiempo de actividad de los usuarios. Ambas características se encuentran descritas en las ecuaciones 2.19 y 2.20 respectivamente.

$$F_{CC}(\nu) = \begin{cases} LN(\mu = 4,15; \sigma^2 = 4,02) & \text{si } \nu < 24MB \\ Pareto(\alpha = 2821,8; \beta = 1,45) & \text{si } 24MB \leq \nu \end{cases} \tag{2.19}$$

$$F_{CC}(\tau) = \begin{cases} Gamma(a = 2,037; b = 2,27; c = 0,30) & \text{si } \tau < 98sec \\ LN(\mu = 10,03; \sigma^2 = 3,42) & \text{si } 98sec \leq \tau < 50min \\ LN(\mu = 9,19; \sigma^2 = 2) & \text{si } 50min \leq \tau \end{cases} \tag{2.20}$$

2.2.6.4. Resumen

A partir del análisis de los diferentes tipos de aplicación existentes para la compartición de ficheros, se pone de manifiesto la complejidad asociada al modelado de este servicio, ya que éste puede estar implementado siguiendo diferentes tipos de enfoques y arquitecturas de comunicaciones.

Asimismo, el tráfico asociado a este tipo de aplicaciones también puede exhibir diferencias a pesar de encontrarse basado en el mismo enfoque y tecnologías. Por ejemplo, en [He et al., 2007] se analizan las características de tráfico de 5 aplicaciones P2P distintas y se concluye que el tráfico de diferentes aplicaciones exhibe diferentes características. Esto hace muy difícil definir un modelo de tráfico que sea válido para cualquier aplicación P2P.

En la literatura existen algunos trabajos con el objetivo de generar o modelar el tráfico de Internet como un conjunto de aplicaciones representativas de Internet. En la tesis doctoral descrita en [Luo, 2005] se propone un modelo de generación de tráfico realista de tráfico de Internet en el que no se tienen en cuenta las aplicaciones de tipo P2P y en su defecto se utilizan aplicaciones FTP. En la tesis doctoral [Svoboda, 2008], se modela el tráfico de Internet en redes inalámbricas en las que tampoco se tienen en cuenta ningún tipo de aplicación de tipo P2P, sólo HTTP, Email y FTP. En definitiva, parece que en la literatura no exista ningún estudio que haga uso de un modelo ampliamente aceptado para modelar el tráfico de aplicaciones de compartición de fichero de tipo P2P. Tal es el caso, en el [Katsaros et al., 2012], donde se describe un generador de tráfico sintético basado en modelos de tráfico de la literatura, donde se afirma que no existe ningún modelo analítico para caracterizar el tamaño de objeto en este tipo de redes. Por estas razones, no existe ningún modelo en la literatura que destaque para caracterizar este tipo de tráfico. En el capítulo 5, se propone un modelo *ad hoc* para caracterizar la demanda de tráfico de las aplicaciones de compartición de ficheros.

2.2.7. Modelos de video sobre Internet

En esta sección, se asocia el video sobre Internet al streaming de video, en el cual, interviene un búfer de datos, que va almacenando el contenido que está siendo descargado, para que la visualización por parte del usuario, sea lo más fluida posible.

2.2.7.1. Conceptos generales

Este servicio de Internet es muy heterogéneo ya que tiene multitud de características que pueden ser diferentes de una aplicación a otra. Por ejemplo, el protocolo de transporte utilizado para el video sobre Internet ha evolucionado con el tiempo. Tradicionalmente

se realizaba mediante protocolos ligeros, como UDP, junto con otros protocolos, como Real Time Streaming Protocol (Protocolo de Flujo en Tiempo Real) (RTSP). En la actualidad, existen infinidad de aplicaciones que hacen uso del protocolo TCP y HTTP. En cuanto a la codificación, a lo largo de los años se han ido utilizado distintos códecs de video. Los códecs que predominan en este tipo de aplicaciones son H.264 y VP8. La elección del códec de video afecta directamente a la tasa de bit necesaria para la transmisión de video y su correspondiente calidad.

En referencia a la distribución del contenido a través de Internet, en las últimas dos décadas de Internet, este servicio de Internet ha evolucionado considerablemente respecto a los enfoques utilizados [Li et al., 2013]:

- **Streaming Cliente-Servidor.** En la primera década a partir de los años 90, el streaming de video era realizado principalmente mediante una arquitectura de cliente-servidor. En ésta época, aparecen los primeros protocolos de streaming, como por ejemplo Real-time Transport Protocol (Protocolo de Transporte de Tiempo real) (RTP).

- **P2P Streaming.** Con objeto de ganar escalabilidad en el número creciente de usuarios del servicio, las tecnologías basadas en P2P para realizar streaming de video han sido analizadas en la última época [Liu et al., 2008]. Mediante este tipo de arquitectura, los clientes del servicio, denominados *peers*, también se convierten en servidores, pues sirven el contenido a otros nodos. Un ejemplo de éxito de este enfoque se encuentra en la aplicación *PPLive*, la cual ha servido miles de videos a millones de usuarios. El principal inconveniente de este enfoque es que los usuarios necesitan instalar una aplicación específica.

- **HTTP Streaming.** En los últimos años, el streaming de video a través de HTTP ha cobrado especial protagonismo, seguramente debido a que los clientes ahora no necesitan descargarse e instalar ninguna aplicación de terceros. Mediante este enfoque, los clientes acceden al contenido multimedia a través de un navegador web. El flujo de video se divide en una secuencia de trozos, denominados *chunks*, los cuales son descargados mediante HTTP. La popularidad de esta estrategia también se debe a la aparición de plataformas de computación en la nube, habilitando que existan servidores de streaming en la nube.

De forma análoga a los servicios de compartición de ficheros, los videos distribuidos a través de Internet pueden tener multitud de características distintas que pueden depender del tipo de contenido o incluso de la plataforma donde se visualiza. Algunos parámetros como la duración, la tasa de codificación y la popularidad dependen íntimamente de la plataforma y el tipo de contenido. Por esta razón, la caracterización de estos parámetros constituye un desafío similar al analizado para los servicios de compartición de ficheros.

A pesar de la heterogeneidad de este servicio, el servicio de streaming de video puede descomponerse en dos fases bien definidas [Rao et al., 2011]: buffering y estado estacionario. En la primera fase de buffering, el cliente empieza a descargar pequeñas porciones de video mientras va llenando un buffer que permita la visualización continuada del video. En la segunda fase, se alcanza un estado estacionario, en el cual se presupone que el cliente consume porciones de video a un ritmo similar que va descargando nuevos *chunks* de video. Este proceso se muestra en la figura 2.9.

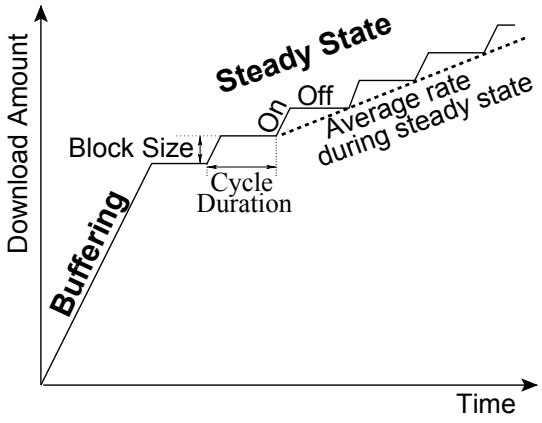

Figura 2.9: Fases del servicio de streaming de video [Rao et al., 2011]

A partir de esta abstracción conceptual, descrita en [Rao et al., 2011], se pueden definir 3 estrategias diferentes para abordar la descarga de *chunks* de video:

1. **Sin ciclos ON-OFF:** todos los datos de video se transfieren durante la fase de buffering, por lo que no existe una fase estable. La ventaja de esta estrategia es que no se necesitan mecanismos complejos en los clientes y servidores, sino que la sesión de streaming de video puede considerarse una simple transferencia de fichero.

2. **Ciclos ON-OFF cortos:** el streaming de video se produce mediante transferencias de bloques de video menores a 2,5 MB, que se corresponde al periodo de actividad (ON). Posteriormente, existe un periodo de inactividad (OFF). Esta estrategia tiene como objetivo mantener una acumulación de contenido en buffer pequeña. Se observarán periodos OFF, cuando la tasa de transferencia de datos sea menor que el el ancho de banda extremo a extremo.

3. **Ciclos ON-OFF largos:** en esta estrategia se transfieren bloques de datos mayores a 2,5 MB, por lo que los ciclos ON-OFF suelen ser mucho mayores comparados con la estrategia anterior. Puede considerarse una estrategia híbrida que busca un compromiso entre el primer y segundo enfoque.

Parámetro	Distribución	Parámetros de ajuste	
Tiempo de actividad de sesión (ON)	Lognormal	$\mu = 5,19$	$\sigma = 1,44$
Tiempo de Inactividad de sesión (OFF)	Exponencial	$\lambda = 5{,}025 \cdot 10^{-6}$	-
Número de transferencias por sesión	Pareto	$\alpha = 1,43$	$\beta = 0,62$
IAT de transferencias de sesión	Lognormal	$\mu = 4,93$	$\sigma = 1,26$
Duración de transferencias	Lognormal	$\mu = 4,29$	$\sigma = 1,28$

Tabla 2.11: Parámetros del modelo de tráfico de video de [Veloso et al., 2002]

A partir del análisis anterior de las características generales del servicio de streaming de video queda claro que existen innumerables configuraciones posibles para aplicaciones de este tipo. Por este motivo, a continuación se analizan diversos modelos de streaming de video sobre Internet, intentando que se cumpla la premisa de que sea lo suficientemente general para que sea representativo para la gran mayoría de aplicaciones existentes de este servicio en Internet. Intentar abordar la caracterización del tráfico de red para cada tipo de aplicación de este servicio (diferentes estrategias de distribución, códecs de video, etc.) queda fuera del alcance de esta tesis doctoral.

2.2.7.2. Modelos generales de streaming de video

Uno de los primeros modelos de tráfico de aplicaciones de streaming de video se encuentra descrito en [Veloso et al., 2002]. En este artículo se presenta una caracterización del tráfico de streaming de video en vivo basándose en un modelo compuesto de 3 niveles: cliente, sesión y transferencia. A nivel de cliente, algunos de los parámetros considerados en este estudio son el número de clientes a lo largo del tiempo, el tiempo entre llegadas de usuarios y la frecuencia de acceso a contenidos. A nivel de sesión, los autores definen un periodo de actividad como el tiempo en el que el cliente se encuentra pidiendo o recibiendo objetos de video y un periodo de inactividad definido a partir de un periodo umbral de inactividad. El nivel de transferencia corresponde a la caracterización más precisa de los periodos de actividad (ON) del nivel superior de sesión. Durante los periodos de actividad a nivel de transferencia, el usuario puede estar descargando uno o más objetos de video. El periodo de inactividad (OFF) se corresponde por tanto a los periodos, durante los cuales el cliente no recibe ningún objeto de video.En este nivel, también se caracterizan la duración de las transferencias de objetos de video, los tiempos entre llegadas de transferencias y número de transferencias simultáneas. En la tabla 2.11 se muestra un resumen de los parámetros más relevantes para la generación de trazas de tráfico para aplicaciones de streaming de video. En este estudio no se consideran aspectos relacionados con la situación de la red, como por ejemplo, ante un escenario con pérdidas o congestión. Destaca que en este estudio no se caracteriza el volumen de datos transferidos durante las sesiones.

En [Tang et al., 2003] se describe un generador sintético de tráfico de streaming de video basado en diferentes modelos para las diferentes características del contenido y del servicio de streaming. Algunas de las características modeladas incluyen la duración de los ficheros, la popularidad del contenido, el proceso de introducción de nuevos ficheros y los patrones de acceso diarios. Este estudio, propone la clasificación de los contenidos en diferentes grupos, los cuales se caracterizan por separado. La duración de los contenidos se modela mediante distribuciones normales, mientras que no se propone ningún modelo para caracterizar las tasas de bit de codificación de los videos. En este estudio también se proponen modelos y distribuciones para caracterizar otras características del servicio, como por ejemplo, la popularidad de los contenidos mediante distribuciones zipf, la aparición de nuevos contenidos a lo largo del día, modelados mediante una distribución de Pareto y la esperanza de vida de los contenidos mediante distribuciones lognormal y Pareto.

En [Costa et al., 2004] se realiza un extenso análisis de trazas de streaming de video, teniendo especial interés en el comportamiento interactivo de los clientes. Debido a las diferencias en las características de las trazas en función del contenido de las mismas, se realiza una clasificación en varias categorías: educacional, entretenimiento de video y entretenimiento de audio. Las principales contribuciones del estudio son la caracterización de los tamaños de ficheros de contenidos y los tiempos entre llegadas de sesiones con diferentes distribuciones en función de la categoría del contenido. Además, de especial interés para esta tesis doctoral, es la caracterización de tráfico en una sesión a partir de los tiempos de actividad (ON) e inactividad (OFF). Para el caso particular de contenidos de video de entretenimiento, se hace una distinción entre videos cortos (de hasta 5 minutos de duración) y video largos. La duración de los tiempos de actividad se modela mediante distribuciones de Pareto y Weibull para videos cortos y largos respectivamente. Los tiempos de inactividad se modelan mediante distribuciones Weibull. A pesar de que se proponen distribuciones para ajustar estas características, este estudio sólo aporta rangos numéricos para ajustar las distribuciones.

En [Hassan et al., 2005] se proponen un conjunto de fuentes de tráfico de aplicaciones multimedia para modelar el tráfico de Internet. Se describen fuentes para las aplicaciones de VoIP, Video y web. En este estudio, al igual que esta tesis doctoral, se hace uso de modelos de fuente de tráfico de tipo ON/OFF para describir tráfico agregado complejo. El tráfico de video lo describen para tres codificaciones diferentes (MPEG-1, MPEG-2 y MPEG-4) y basándose en un modelo $M/G/\infty$ con los siguientes parámetros de entrada: escala de tiempos o nivel de Group Of Pictures (Grupo De Imágenes) (GOP), distribución de tamaño de frame o de GOP, tipo de correlación (LRD o SRD) y distribución de tamaño de paquetes.

En [Forconi et al., 2008] se propone un modelo para la generación de tráfico de

Componente	Distribución	Parámetros	Estadísticos
IAT entre frames	Determinista	100 ms (basado en 10 fps)	-
Número de paquetes por frame	Determinista	8 paquetes por frame	-
Tamaño de paquetes	Pareto truncada	$\alpha = 1,2$ $k = 40bytes$ $max = 250bytes$	$E[X] = 100bytes$ $Max = 250bytes$
IAT entre paquetes	Pareto truncada	$\alpha = 1,2$ $k = 2,5$ $max = 12,5$	$E[X] = 6ms$ $Max = 12,5$

Tabla 2.12: Modelo de tráfico de streaming de video de [Srinivasan et al., 2008]

Internet a partir de varios modelos para diferentes aplicaciones. El modelo de aplicación de streaming de video utilizada en este estudio, se corresponde a la descrita en [Srinivasan et al., 2008]. En la tabla 2.12 se muestra como este modelo ajusta el tamaño de paquetes y su tiempo entre llegadas mediante distribuciones de Pareto truncadas. Este modelo se encuentra ajustado para un servicio de video de streaming con una tasa de fuente de 64 Kbps. Este hecho hace que este modelo de tráfico de video sea poco realista con las tasas de bit de los servicios de video actuales.

En [García et al., 2007] se presenta una extensa caracterización del servicio de video bajo demanda a partir de las trazas recogidas durante más de 4 años para videos de la red *RealNetworks*. Los autores presentan un estudio estadístico sobre el comportamiento de usuario y el tráfico de streaming generado, con el principal objetivo de aportar la información necesaria para poder desarrollar modelos de simulación y de generación de tráfico que puedan ser utilizados en diferentes escenarios y bajo diferentes condiciones de servicio. El estudio realiza dos análisis bien diferenciados sobre el comportamiento de usuario y sobre la caracterización del tráfico. El análisis de comportamiento de usuario se estudia mediante diversas características de las sesiones. El número de reproducciones por sesión se modela mediante dos distribuciones zipf-like con parámetros $\theta = 1,77$ para el 94,28 % de los casos que tienen de 1 a 6 reproducciones, y $\theta = 3,09$ para el %5,72 que tienen de 6 a 25 reproducciones. El tiempo entre reproducciones se ajusta a una distribución de tipo Weibull con parámetros $\alpha = 0,16687$ y $\beta = 0,51107$. Además, también caracterizan la duración de los videos reproducidos, haciendo una distinción entre videos largos y cortos. Los autores caracterizan ambos videos con una concatenación de distribuciones exponenciales con parámetros $\mu_1 = 0,16$ y $\mu_2 = 0,06$ para videos cortos, y $\mu_1 = 0,2$ y $\mu_2 = 0,27$ para videos largos. Además, este estudio también analiza las interrupciones, los saltos en la visualización y la popularidad. En cuanto a la

	QL1	QL2	QL3	QL4	QL5
Tasa de bit	1920 Kbps	960 Kbps	480 Kbps	240 Kbps	120 Kbps
Tasa de frame			25 fps		
Distribución de Pareto	$\alpha = 1,2$ $k = 4800$ $m = 26100$	$\alpha = 1,2$ $k = 2400$ $m = 13100$	$\alpha = 1,2$ $k = 1200$ $m = 6500$	$\alpha = 1,2$ $k = 600$ $m = 3300$	$\alpha = 1,2$ $k = 300$ $m = 1654$

Tabla 2.13: Modelo de tráfico de streaming de video de [Zou et al., 2013]

caracterización del tráfico de video, los autores modelan el tamaño de los de paquetes de video y audio, y su tiempo entre llegadas. No obstante, es importante destacar que la caracterización del tráfico puede no ser muy representativa debido a la baja calidad de las trazas analizadas, con calidades de video de 140 Kbps a 20 fps y 90 Kbps a 15 fps. Además, cabe mencionar que el códec de video utilizado, *RealVideo*, es un códec propietario basado en H.263.

En [Zou et al., 2013] se propone un esquema adaptativo y orientado a dispositivos para sistemas Long Term Evolution (LTE) con el objetivo de hacer un uso eficiente de los recursos de red y proveer de Quality of Experience (Calidad de Experiencia) (QoE) a los usuarios finales de un servicio de distribución de contenidos multimedia. Para analizar el rendimiento de este esquema se realizan simulaciones de servicios de streaming de video utilizando un modelo de tráfico basado en distribuciones de Pareto truncadas, derivado del estudio descrito en [Srinivasan et al., 2008]. En este estudio se ajusta el modelo de tráfico de video para diferentes tasas de bit (Tabla 2.13), donde se expresan los parámetros de la función de Pareto (forma α y escala k) así como el valor máximo (m).

2.2.7.3. Modelos de streaming de video sobre HTTP

Similar a los modelos vistos para la aplicación web, en [Chen et al., 2008] se propone un modelo de tráfico de video sobre HTTP, donde no sólo se tiene en cuenta el tráfico generado durante la transmisión de video, sino todo el proceso a partir del cual el usuario llega a ese contenido. En la figura 2.10 se muestran los 4 niveles en los que se compone el modelo descrito: sesión, página, objeto y paquete. En este estudio, se ajustan los parámetros del modelo de tráfico mediante las trazas de tráfico obtenidas a lo largo de 10 días, para 42 usuarios de una red local. En la tabla 2.14 se muestra cada uno de los parámetros de los cuatro niveles en los que se basa el modelo, junto con las distribuciones propuestas y sus estadísticos (media y desviación típica).

En [Zink et al., 2009] se presenta un estudio de *YouTube*, el servicio web de video online más popular en la red. Las medidas en las que se basa el estudio se han realizado

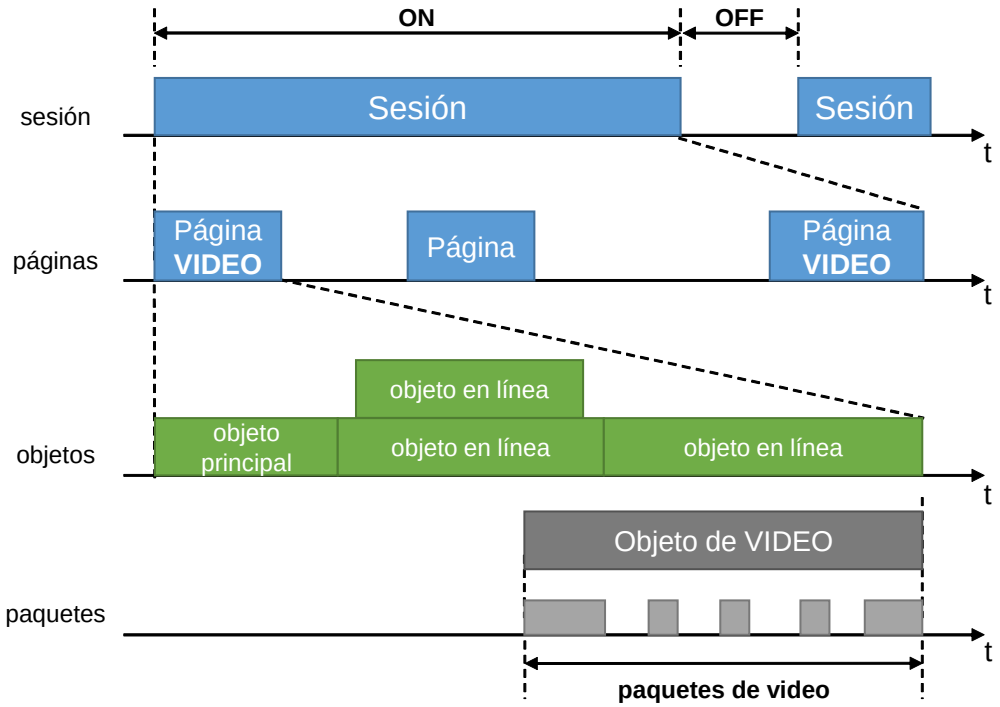

Figura 2.10: Modelo de página de video sobre HTTP de [Chen et al., 2008]

a lo largo de casi un año en un campus universitario, analizando algunas características del servicio, entre las que destacan la duración, la tasa de datos de las sesiones de streaming y la popularidad de los videos. De entre los datos más interesantes extraídos se encuentran algunas estadísticas sobre las tasas de bit de los videos visualizados por los usuarios del campus, cuyas medias oscilan entre 632 Kbps y 908 Kbps.

De forma similar al anterior estudio, en [Abhari and Soraya, 2010] se introduce una caracterización de tráfico de red del servicio de streaming de video de *YouTube*. A partir de trazas de tráfico capturas, este estudio propone un modelo para la generación de tráfico a partir de las características y distribuciones descritas en la tabla 2.15. Además, también se analizan otros parámetros, como por ejemplo, la valoración y el número de visualizaciones de los videos.

En [Rao et al., 2011] se presenta un estudio de las características de red de los dos servicios de video más populares: *YouTube* y *Netflix*. Se identifican 3 estrategias diferentes para el streaming de video que producen patrones de tráfico distintos: sin ciclos ON-OFF, ciclos ON-OFF cortos y ciclos ON-OFF largos. Además, los autores definen un modelo para streaming de video con y sin interrupciones, aunque sólo desde un punto de vista teórico.

Parámetro	Media	Desv. típica	Distribución
IAT sesión	101,18 s	128,3 s	Weibull ($\alpha = 88,9448; \beta = 0,7956$)
Número de paquete de llamada	242,99	415,83	Normal ($\mu = 242,99; \sigma = 415,83$)
Proporción de página de video	0,1185	0,2011	Weibull ($\alpha = 0,0817; \beta = 0,6175$)
Elementos en línea por página	4,2623	8,3995	Lognormal ($\mu = 0,6569; \sigma = 1,2593$)
Tiempo de lectura	8,7564 s	8,7564 s	Exponential ($\mu = 8,7564$)
Tamaño de objeto principal	15,548 KB	86,032 KB	Lognormal ($\mu = 7,9249; \sigma = 1,8584$)
Tamaño de objeto en línea	99,101 KB	270,90 KB	Weibull ($\alpha = 37,217; \beta = 0,4362$)
Tiempo de parseo	0,8154 s	1,4950 s	Weibull ($\alpha = 0,5181; \beta = 0,5803$)
Tamaño de objeto de video	9,1841 MB	7,6874 MB	Weibull ($\alpha = 9,7631; \beta = 1,1998$)
Tiempo de llegada relativo de obj. video	1,7339 sec	4,8948 s	Weibull ($\alpha = 0,6196; \beta = 0,4273$)
IAT de paquetes de video	11,84 ns	57,73 ms	Weibull ($\alpha = 1,5794; \beta = 0,3151$)
Tamaño paquetes de video	1313,23 B	329,23 B	Exponential ($\mu = 329,23$)

Tabla 2.14: Parámetros del modelo de tráfico de video de [Chen et al., 2008]

2.2.7.4. Resumen

De forma análoga al servicio de compartición de ficheros, este servicio engloba diversas aplicaciones de streaming basadas en enfoques y tecnologías muy diversas. Tales son las diferencias en aspectos relevantes (arquitectura de red, estrategia de streaming, códec de video, etc.) que el tráfico de red generado por diferentes aplicaciones probablemente tengan características muy distintas entre sí. Quizás la aplicación de streaming de video más relevante en la actualidad sea el streaming de video a través de HTTP. No obstante, aún dentro de esta categoría existen muchos enfoques distintos que hacen variar el tipo de tráfico generado.

A partir del análisis de los modelos de la literatura, se aprecia como muchos modelos

Parámetro	Media	Distribución
Duración del video	4,55	Log-Logistica ($a = 2,54; b = 226,54$)
Tamaño de fichero	9809,52 KB	Gamma ($a = 1,8; 5441,94$)
Número de visualizaciones	541564,31	Weibull ($\alpha = 0,55; \beta = 290130$)
Número de valoraciones	1309	Weibull ($\alpha = 0,52; \beta = 629,24$)
Valoración media	4,56	Weibull ($\alpha = 15,95; \beta = 4,73$)

Tabla 2.15: Caracterización de los videos de *YouTube* en [Abhari and Soraya, 2010]

sólo hacen referencia a características temporales, es decir, caracterizan el servicio mediante el análisis de los periodos de actividad e inactividad a nivel de sesión, como por ejemplo en el modelo de [Veloso et al., 2002]. Según este modelo, un usuario de este servicio no utiliza todo su ancho de banda durante este tiempo de actividad, sino que va descargando fragmentos del video a la vez que los va visionando mediante el uso de buffers. Por su nivel de abstracción, este tipo de modelos no sirve para caracterizar la carga de tráfico de un usuario de este servicio.

En otros modelos, se aprecia que han sido generado a partir de trazas de red con unas condiciones determinadas y para un servicio con una calidad específica, como por ejemplo es el caso de los modelos expuestos en [García et al., 2007, Srinivasan et al., 2008]. Este hecho hace que tanto la caracterización de la duración de los periodos de actividad, como la de los tamaños de los paquetes o frames, dependan íntimamente de las condiciones de red en las que las trazas fueron capturadas (ancho de banda disponible por cliente, congestión, etc.).

El modelo propuesto por [Chen et al., 2008] parece ser el más adecuado para caracterizar la demanda de tráfico que realizan los usuarios de este servicio. No obstante, el modelo cuenta con dos importantes desventajas: su alta complejidad y su escasa validación (modelo ajustado a partir de trazas de 42 usuarios).

No obstante, esta tesis hace uso del modelo de generación de tráfico propuesto en [Zou et al., 2013], puesto que cuenta con una gran flexibilidad a la hora de seleccionar la calidad de video de streaming mediante la tasas de bit de codificación. Además, hay que destacar que fue desarrollado para el mismo fin que esta tesis doctoral, estimar la demanda de tráfico para analizar el rendimiento de redes de comunicaciones.

2.2.8. Modelos de juegos en red

El último de los servicios dentro del conjunto de servicios representativos de Internet corresponde a los juegos en red. A pesar de no ser un servicio que tenga grandes requisitos en cuanto ancho de banda se refiere, su popularización en las últimas décadas ha hecho que una parte no despreciable del tráfico global de Internet se deba a aplicaciones de este tipo [Ratti et al., 2010].

2.2.8.1. Conceptos básicos

Los videojuegos llevan siendo parte de la vida de las personas desde hace ya varias décadas, desde la primera aparición del famoso *Pong*, hasta el desarrollo de plataformas de juegos dedicadas (por ejemplo, las consolas como *Nintendo* o *Playstation).* Los videojuegos en red comienzan su popularidad a principios de los años 90, cuando *idSoftware* desarrolla un juego llamado *Doom* [Armitage et al., 2006]. A pesar de no ser el primer juego de First Person Shooter (Disparos en Primer Persona) (FPS), fue el

primero con la funcionalidad de juego en red. El gran éxito cosechado con esta nueva funcionalidad sentó las bases de los videojuegos en red, hasta el punto de la existencia de muchos de ellos que sólo pueden jugarse online. A partir de entonces han surgido infinidad de juegos en red, pudiéndose clasificarse en función sus características:

- **En primera persona.** En este tipo de juego los jugadores luchan unos contra otros, estando o no separados en equipos, y con un objetivo para ganar la partida y/o ronda. Los juegos más conocidos de esta categoría se conocen como juegos de FPS, como por ejemplo, *Counter Strike*, *Quake*, *Battlefield*, etc. La arquitectura más habitual en este tipo de juegos suele ser de tipo cliente-servidor.

- **Estrategia en tiempo real.** En este tipo de juegos, los jugadores suelen compartir un mismo escenario o mapa. La interacción entre jugadores no suele ser tan frenética como en los juegos en primera persona ya que en el juego prima la estrategia. Un juego representativo de esta categoría es *Starcraft*, en el cual los jugadores comparten un mismo mapa en el que tienen que ir construyendo edificios y unidades de forma que la estrategia final sea acabar con las tropas de los oponentes. La arquitectura de este tipo de juegos puede ser muy variada, aunque predominan las basadas en P2P.

- **Multi-jugador masivos en línea.** Este tipo de juegos se le denomina Massively Multiplayer Online Game (MMOG). Un caso particular de este tipo de videojuegos también son los conocidos como Massively Multiplayer Online Role-Playing Game (MMORPG). Los jugadores de este tipo de videojuegos disponen de un personaje que juega en un mundo virtual, que a diferencia de los anteriores tipos de videojuegos, es mucho más grande. El objetivo de este tipo de juegos suele ser la de completar misiones designadas en el juego. Habitualmente, a medida que el jugador gana experiencia y completa misiones su personaje sube de nivel, lo cual hace que su personaje gane habilidades y se haga más poderoso. Debido a la gran cantidad de usuarios conectados simultáneamente, la arquitectura más habitual se basa en clústers cliente-servidor. Uno de los juegos más representativos es el World of Warcraft (WoW).

- **Casuales.** En esta categoría se encuentran los juegos online de uso esporádico, habitualmente disponibles en páginas web o en redes sociales. En general, son juegos que son independientes de la plataforma, ya que se encuentran basados en tecnologías web. Este tipo de juegos ha adquirido cierta popularidad en los últimos años y el tráfico que generan, se encuentra enmascarado con el tráfico de navegación web.

En relación con la última categoría de juegos online, seguramente debido a su naturaleza esporádica, y que se encuentran dentro de plataformas web o redes sociales,

en la literatura no se encuentran contribuciones relevantes para la caracterización del tráfico que generan. Por este motivo en esta tesis doctoral, se prestará mayor atención a las otras categorías.

Arquitectura de red. Por su propia naturaleza, los videojuegos suelen involucrar un conjunto elevado de acciones que requieren ciertas características de la red de comunicaciones. Por ejemplo, la gran mayoría de videojuegos necesitan respuestas rápidas (baja latencia), pocas pérdidas de paquetes y mucha periodicidad en el intercambio de información entre clientes y servidores (en caso de haberlos).

La comunicación entre nodos de un juego online puede basarse en diferentes tipos de arquitecturas de red:

- ***Peer-to-Peer.*** El primer tipo de arquitectura en los juegos de red es la de *Peer-to-Peer*, pues todos los nodos se comunicaban entre sí sin necesidad de un servidor. Este tipo de arquitectura es bastante usual en los juegos de estrategia en tiempo real. Por ejemplo, *Starcraft* se encuentra basado en una arquitectura P2P, en la que los jugadores intercambian entre sí la información de la partida. No obstante, *Starcraft* cuenta con servidores centrales que sirven para que los jugadores encuentren a otros jugadores e inicien la partida. El tráfico entre *peers* puede ser muy variado, aunque tiende a tener una alta periodicidad.

- **Cliente-Servidor.** Este tipo de arquitecturas tiene una mejor escalabilidad de jugadores que la anterior. Todos los clientes envían las actualizaciones del juego a un servidor central. El servidor central procesa toda la información recibida y transmite los resultados a todos los jugadores. Muchos juegos de primera persona se encuentran basados en esta arquitectura. El tráfico de servidor a clientes tiende a ser a ráfagas, ya que el servidor envía información del estado de todos los jugadores a cada cliente. En cuanto el tráfico de clientes a servidor suele estar caracterizado por una tasa de bit constante.

- **Clúster Cliente-Servidor.** Cuando la población de jugadores es realmente alta, con miles e incluso cientos de miles de jugadores simultáneos se encuentran conectados al mismo juego en red, es necesario aumentar la escalabilidad mediante el uso de una arquitectura basada en clústers cliente-servidor. De esta forma, los recursos de red y computacionales pueden ser divididos en varios servidores. Este es el caso de los juegos correspondientes a la categoría MMOG, como por ejemplo el popular juego ya mencionado, WoW. El tráfico entre clientes y servidores depende íntimamente del juego y de las acciones en el mismo. En líneas generales, el tráfico también muestra una muy alta periodicidad y el tráfico servidor-cliente suele ser más intenso que en la dirección opuesta.

En la figura 2.11 se muestran las 3 arquitecturas más relevantes en la actualidad [Svoboda, 2008].

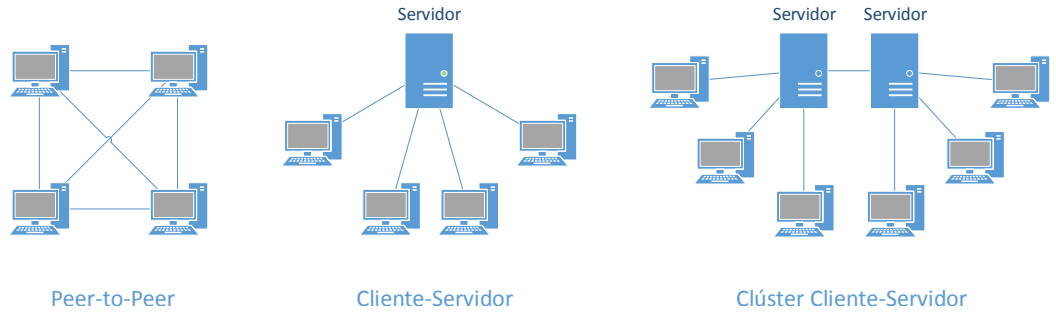

Figura 2.11: Arquitecturas de comunicaciones típicas de juegos en red

2.2.8.2. Modelos de juegos en primera persona

El artículo descrito en [Borella, 2000] contribuye a la caracterización del tráfico observado durante el uso del juego FPS *Quake*. Se describen dos modelos para generar tráfico de este tipo a partir de las trazas observadas durante dos partidas del juego con diferente número de usuarios conectados. En la comunicación cliente-servidor, se propone un modelo basado en tamaños de paquetes fijos con valor de 24 bytes y en distribuciones de valor extremo para el tiempo entre llegada de paquetes que difiere en función del cliente analizado. La comunicación del servidor hacia los clientes se modela con distribuciones de valor extremo, tanto para el tamaño de paquete como para el tiempo entre llegadas. Los valores propuestos para las distribuciones varían para cada traza y cliente observado, por lo que no existe un modelo analítico general.

Uno de los primeros modelos analíticos de tráfico de red para juegos FPS se encuentra en [Färber, 2002]. Este artículo presenta la caracterización de tráfico del popular *Counter Strike* basada en la captura de tráfico en una red y observando los diferentes patrones existentes durante la diferentes fases del juego. Destaca las grandes semejanzas que comparte este modelo con otros juegos del mismo tipo, como es el caso del Unreal Tournament. En la tabla 2.16 se muestra el modelo de tráfico de red propuesto por el autor, el cual se basa en la distribución de valor extremo para todos sus parámetros. El sentido del tráfico se encuentra expresado desde el punto de vista de los clientes.

En [Lang and Armitage, 2003] se presenta un modelo de simulación de tráfico para el juego de la videoconsola *Xbox* llamado *Halo*. Se analizan algunos parámetros de tráfico de red como la longitud de los paquetes, su tasa y el tiempo entre llegadas de los mismos. En la tabla 2.17 se muestran los valores de los parámetros del modelo. Destaca que los tamaños de los paquetes, tanto en sentido de subida como de bajada de cliente, depende del número de clientes conectados. Los IATs de paquetes son, en la mayoría de

Parámetro del Modelo	**Distribución**
Tamaño de Paquete (Bajada)	Valor Extremo ($a = 55; b = 6ms$)
Tamaño de Paquete (Subida)	Determinista ($a = 40ms$)
IAT (Bajada)	Valor Extremo ($a = 120; b = 36bytes$)
IAT (Subida)	Valor Extremo ($a = 80; b = 5,7bytes$)

Tabla 2.16: Modelo de tráfico de juego *Counter Strike* de [Färber, 2002]

Parámetro del Modelo	Probabilidad	**Distribución**
Tamaño de Paquete (Bajada)	-	$Longitud = 30 * N_clientes + 100$
Tamaño de Paquete (Subida)	0,16	Determinista ($a = 72bytes$)
	0,84	$Longitud = 30 * N_clientes + 80$
IAT (Bajada)	-	Determinista ($a = 40ms$)
IAT (Subida)	0,67	Determinista ($a = 40ms$)
	0,33	Uniforme ($a = 0; b = 40ms$)

Tabla 2.17: Modelo de tráfico de juego *Halo* de [Lang and Armitage, 2003]

los casos, de un valor igual o inferior a 40 milisegundos.

En [Lang et al., 2003] se realiza un estudio sobre el popular juego FPS *Half-Life*. A partir de trazas de red del juego, se caracteriza el tráfico de red entre servidor y clientes. En sentido servidor-cliente, el IAT es de aproximadamente 60 ms la mitad de las veces y 70 ms la otra mitad. En el caso del tamaño de paquete, los resultados no son concluyentes, pues a pesar de afirmar que dependen del mapa del juego y que se ajustan a una distribución lognormal, no aportan un ajuste mediante parámetros. El IAT entre paquetes cliente-servidor depende de cómo se renderizan los gráficos (mediante *OpenGL*, *Direct3D* o software), aunque se encuentren próximos a los 40 ms. El tamaño de paquetes cliente-servidor se ajusta mediante una distribución normal con media 72,29 y desviación típica de 6,97 bytes.

Un estudio muy similar al anterior se encuentra en [Lang et al., 2004]. Este artículo presenta los resultados del desarrollo de un modelo sintético para generar tráfico para el popular juego de tipo FSP Quake3, generado a partir de trazas de red. El IAT servidor-clientes se modela de forma determinista con un valor en 50 ms. El tamaño de paquete del servidor a clientes se modela mediante una distribución lognormal ($\mu = 79,34; \sigma = 0,25$ bytes) más un incremento por cada jugador superior a 2 que viene dado por una distribución exponencial ($\mu = 13bytes$). El IAT de los clientes vuelve a depender de factores externos, como es en este caso, de las tarjetas gráficas de los jugadores. El tamaño de paquetes se modela mediante una distribución normal ($\mu = 64,15; \sigma = 3,20$).

Otro estudio de algunos de los autores anteriores [Zander and Armitage, 2005],

Parámetro del Modelo	**Distribución**
Tamaño de Paquete (Servidor)	Lognormal ($\mu = 4,21; \sigma = 0,657$)
Tamaño de Paquete (Cliente)	Normal ($\mu = 73,53; \sigma = 0,2331$)
IAT (Servidor)	Normal ($\mu = 31,23; \sigma = 1,221$)
IAT bajo (Cliente)	Normal ($\mu = 32,70; \sigma = 8,739$)
IAT alto (Cliente)	Normal ($\mu = 218,50; \sigma = 15,130$)

Tabla 2.18: Modelo de tráfico de juego *Unreal Tournament 99* de [Svoboda, 2008]

describe las características del tráfico generado por el juego de tipo FPS, *Halo 2*. El principal objetivo que se persigue es la de proveer la información necesaria para estimar la cantidad de tráfico que genera y consume el juego, y su consecuente impacto en las redes de acceso. En este estudio se presentan datos sobre la tasa de paquetes, ancho de banda, tamaño de paquetes y tiempo entre llegadas. El tamaño de paquetes en ambos sentidos se modelan mediante distribuciones de valor extremo, cuyos parámetros aumentan en función del número de usuarios conectados. El IAT de paquetes cliente-servidor se ajusta a una distribución normal ($\mu = 40; \sigma = 1ms$). El IAT de paquetes servidor-cliente se modela mediante una distribución de valor extremo ($a = 39,7; b = 1,9ms$).

En [Wattimena, 2006] se presenta un informe técnico que tiene como principal objetivo estudiar la calidad de experiencia de los jugadores así como los tiempos de respuesta que experimentan. Este estudio ofrece algunas recomendaciones dirigidas a los administradores de los servidores de juegos para proveer de mayor calidad de experiencia a sus jugadores. En cuanto a los parámetros utilizados para caracterizar los tráficos de subida y de bajada de los clientes con el servidor de juegos, utilizan valores y características muy similares a las vistas en otros modelos de la literatura, con tiempos de llegada entre paquetes de 40 ms y longitudes de 125 bytes en sentido descendente y 80 bytes en sentido ascendente (de clientes a servidor).

En [Svoboda, 2008] se proponen modelos de tráfico de red para un juego de cada una de las categorías anteriormente descritas, a excepción de los juegos casuales. Para el caso de los juegos en primera persona, el juego representativo escogido por el autor ha sido el *Unreal Tournament 99*. EL modelo de tráfico extraído se basa en las capturas de red realizadas en un ordenador conectado a Internet. En la tabla 2.18 se muestran un resumen de los resultados de interés extraídos para el juego.

En [Asensio et al., 2008] se realiza un análisis del ancho de banda necesario para juegos multigador en línea, correspondientes en su gran mayoría a esta categoría. A pesar de que no se proponen distribuciones para caracterizar el tráfico de red de ningún juego, se aportan datos sobre el ancho de banda necesario tanto en dirección cliente-servidor como al revés. Particularizando en juegos FPS, el estudio estima que cada cliente requiere de un ancho de banda mínimo de 85 Kbps de bajada para el juego *Counter Strike*, y

Parámetro del Modelo	Distribución
Tamaño de Paquete (Bajada)	Valor Extremo ($a = 50; b = 4,5ms$)
Tamaño de Paquete (Subida)	Normal ($a = 40; b = 6ms$)
IAT (Bajada)	Valor Extremo ($a = 330; b = 82bytes$)
IAT (Subida)	Valor Extremo ($a = 45; b = 5,7bytes$)

Tabla 2.19: Modelo de tráfico de juegos FPS de [Srinivasan et al., 2008]

200 Kbps para el juego *Quake III Arena*. El ancho de banda de subida es de 47 Kbps y 60 Kbps para los juegos anteriormente mencionados.

En el estudio anteriormente mencionado [Srinivasan et al., 2008], también se describe un modelo para juegos de tipo FPS, ya que afirman que son una buena representación de los requisitos necesarios en los juegos modernos de tipo MMOG. Este estudio propone un modelo basado en distribuciones de valor extremo. En la comunicación, en el enlace descendente, desde el punto de vista del cliente, el valor medio del tamaño de paquetes de esta distribución se encuentra en 330 bytes y el IAT de paquetes en 50 ms. En la tabla 2.19 se muestran los parámetros del modelo, tanto en el enlace de bajada como de subida.

En [Ratti et al., 2010] se presenta un exhaustivo estudio de la literatura existente sobre el tráfico en red de los juegos de disparos en primera persona. Los autores analizan los resultados de otros artículos sobre el tráfico de los juegos en términos de los IATs y tamaños de paquetes, tanto desde el servidor al cliente como en sentido opuesto.

2.2.8.3. Modelos de juegos de estrategia en tiempo real

En [Dainotti et al., 2005] se analiza el tráfico de red generado por el juego *Starcraft* y se realiza una caracterización estadística a nivel de paquete teniendo en cuenta el número de jugadores que pueden estar en una partida. Los autores confirman que el tráfico generado tiene una periodicidad muy alta de paquetes de pequeño tamaño y que los tamaños de paquetes no crecen con el número de jugadores, por lo que este estudio es válido independientemente del número de jugadores conectados. Además, se desarrolla un modelo analítico para aproximar el tráfico observado en términos de paquetes de entrada y salida de los clientes, así como sus correspondientes IATs. En la tabla 2.20 se muestran los resultados del estudio, tanto para el tráfico entrante como saliente desde la perspectiva de los clientes.

Otro modelo de tráfico propuesto en [Svoboda, 2008] es para un juego de estrategia en tiempo real, *Starcraft*. Para el caso de este juego, el autor afirma que los valores para el IAT y el tamaño de paquete es prácticamente constante. El IAT de paquetes oscila entre 31 y 33 ms, mientras que sus tamaños se encuentran entre 75 y 78 bytes.

Parámetro del Modelo	**Distribución**	Probabilidad	Parámetros
Tamaño de Paquete (Entrante)	Determinista	$p = 0,032$	$a = 16$
	Determinista	$p = 0,108$	$a = 17$
	Determinista	$p = 0,724$	$a = 23$
	Determinista	$p = 0,062$	$a = 27$
	Determinista	$p = 0,074$	$a = 33$
Tamaño de Paquete (Saliente)	Determinista	$p = 0,062$	$a = 16$
	Determinista	$p = 0,109$	$a = 17$
	Determinista	$p = 0,742$	$a = 23$
	Determinista	$p = 0,087$	$a = 27$
IAT (Entrante)	Exponencial	-	$\mu = 0,043633$
IAT (Saliente)	Determinista	$p = 0,662$	$a = 0$
	Uniforme	$p = 0,278$	$a = 0,05; b = 0,17$
	Determinista	$p = 0,06$	$a = 21$

Tabla 2.20: Modelo de tráfico de juego *Starcraft* de [Dainotti et al., 2005]

2.2.8.4. Modelos de juegos multijugador masivos en línea

En [Svoboda, 2008] también se propone un modelo de tráfico para el popular juego de tipo MMOG, *WoW*. El autor propone un modelo para el tráfico de red en sentido descendente al cliente y otro en sentido ascendente. En sentido servidor-cliente el tamaño de paquetes se define mediante una distribución Weibull ($\sigma = 426; k = 0,8196$) con un límite superior situado en los 3010 bytes. En sentido ascendente, el tamaño de paquetes se caracteriza con 3 tamaños diferentes mediante una combinación de deltas de Dirác ($a = 6, b = 19, c = 43byte$) (ecuación 2.21). El IAT entre paquetes se caracteriza mediante una combinación de distribuciones uniformes ($a = 218,3; b = 251,2; c = 1500ms$) (ecuación 2.22).

$$f(x; a, b, c) = 0,52 \cdot \delta(x - a) + 0,14 \cdot \delta(x - b) + 0,34 \cdot \delta(x - c) \tag{2.21}$$

$$f(x; a, b, c) = \begin{cases} 0,620 \cdot \frac{1}{a-0} & \text{si } 0 \geq x < a \\ 0,257 \cdot \frac{1}{b-a} & \text{si } a \geq x < b \\ 0,123 \cdot \frac{1}{c-b} & \text{si } b \geq x \leq c \\ 0 & resto \end{cases} \tag{2.22}$$

En [Chen et al., 2005] se presenta un análisis de las trazas de paquetes realizadas a un juego masivo online llamado *ShenZhou Online*. El estudio identifica algunas propiedades particulares de los juegos de este tipo, en términos de tamaños de paquetes, su periodicidad y dependencia temporal en la llegada de paquetes. La longitud de los paquetes generado por los clientes es muy pequeña en comparación con la de los

servidores de este juego. El 98 % de los paquetes de clientes tienen un tamaño igual o menor a 31 bytes. No obstante, la carga útil de los paquetes de servidor (*payload*) tiene una distribución mucho más amplia y con una media de 114 bytes. Este artículo también caracteriza el ancho de banda utilizado en media para cada cliente, destacando el hecho de que se necesita una tasa de bit media de 10 Kbps. Esta característica es mucho menor a las identificadas para otros juegos, como por ejemplo, el anteriormente descrito *Counter Strike* con 40 Kbps. De la misma forma, el tiempo entre llegadas de paquetes también es menos restrictivo que en los juegos de tipo FPS, ya que tienen un valor que oscila entre 0 y 600 ms en comparación con valores cercanos a 40 ms típicos de los juegos en primera persona.

2.2.8.5. Resumen

Anteriormente se han analizado algunos de los modelos de tráfico de juegos en red más relevantes en la literatura. Como conclusión, se puede extraer que los juegos de acción son los más exigentes en cuanto a requisitos de red se refiere. Por esta razón, a la hora de dimensionar las redes de acceso se puede optar por considerar el caso peor de requisitos y condiciones de red mediante el uso de modelos de este tipo de juegos en red.

El modelo descrito en [Srinivasan et al., 2008] constituye uno de los modelos más recientes para los juegos de acción y se encuentra en línea con otros modelos del mismo tipo de juegos [Färber, 2002, Borella, 2000, Zander and Armitage, 2005]. Además, hay que resaltar que los valores de los tamaños de paquetes y tiempo entre llegadas también se corresponden con las medidas del resto de modelos, siendo estos valores muy similares en todos ellos. Por las razones anteriores, esta tesis hace uso del modelo de [Srinivasan et al., 2008] para la caracterización de este tipo de aplicaciones.

2.3. Dimensionado de redes de acceso

Esta sección comienza con un análisis de algunas de las características de rendimiento de una red, como la velocidad y la capacidad. Además, se argumenta porqué estas métricas no son idóneas para medir el rendimiento de una red y se proponen algunas alternativas.

Posteriormente, se analizan las principales tecnologías utilizadas en las redes de acceso a Internet, haciendo hincapié en los diferentes elementos de su arquitectura. El objetivo principal del capítulo reside en caracterizar estas arquitecturas mediante un modelo de referencia de red que sea válido para todas las tecnologías y posibles escenarios. Este esquema de red será utilizado posteriormente en el desarrollo de la metodología de estimación de demanda de tráfico en redes de acceso de Internet, descrita en el capítulo 4 de esta tesis doctoral.

2.3.1. Métricas relativas al ancho de banda de una red

Existen numerosas formas de definir el rendimiento de una red de comunicaciones en referencia al ancho de banda disponible o percibido por el usuario. Por esta razón existe cierta confusión a la hora de utilizar algunos términos como la velocidad o la capacidad de la red para caracterizar la calidad o rendimiento de una red.

2.3.1.1. Velocidad

La velocidad de una red se define como la tasa a la que el tráfico de un usuario es transportado a través de una red en un momento específico del tiempo.

La gran mayoría de proveedores de servicio de Internet ofertan la velocidad máxima que puede experimentar un usuario en una conexión, pero sin llegar a especificar las condiciones para que se llegue a esa cota superior. En general, la velocidad percibida por el usuario en una red de acceso de banda ancha depende íntimamente de algunos parámetros como por ejemplo, la arquitectura o el uso del resto de usuarios de la red.

Es por tanto, que la definición del término velocidad no puede considerarse trivial. Como se ha mencionado anteriormente, la velocidad depende de la demanda que soporta la red en un momento determinado, por lo que una única medida de velocidad no es por sí misma una métrica válida de rendimiento de una red de comunicaciones.

Siguiendo el ejemplo similar al descrito en [Adtran, 2009], la figura 2.12 muestra un ejemplo de 10 medidas de velocidad en una red de acceso. A partir de este ejemplo, se pueden extraer las siguientes medidas:

- El valor máximo es de 22 Mbps. No obstante, ninguna otra medida de velocidad se aproxima tan siquiera a este valor.
- La media aritmética de las muestras es de 6 Mbps. Sin embargo, sólo el 30 % de las muestras superan este valor.
- La mediana es de 3 Mbps, es decir, la mitad de las muestras exceden este valor. En otras palabras, se podría esperar superar esta velocidad con una probabilidad del 50 %.

Este ejemplo confirma la afirmación anterior de que la velocidad no es una métrica válida para cuantificar el rendimiento de una red a lo largo de un periodo de tiempo.

2.3.1.2. Capacidad

A diferencia del término de velocidad, la capacidad tiene una definición mucho más precisa. La capacidad de una red es un valor constante que depende únicamente del diseño de la propia red y no de las demandas de tráfico de los usuarios. Se define, por

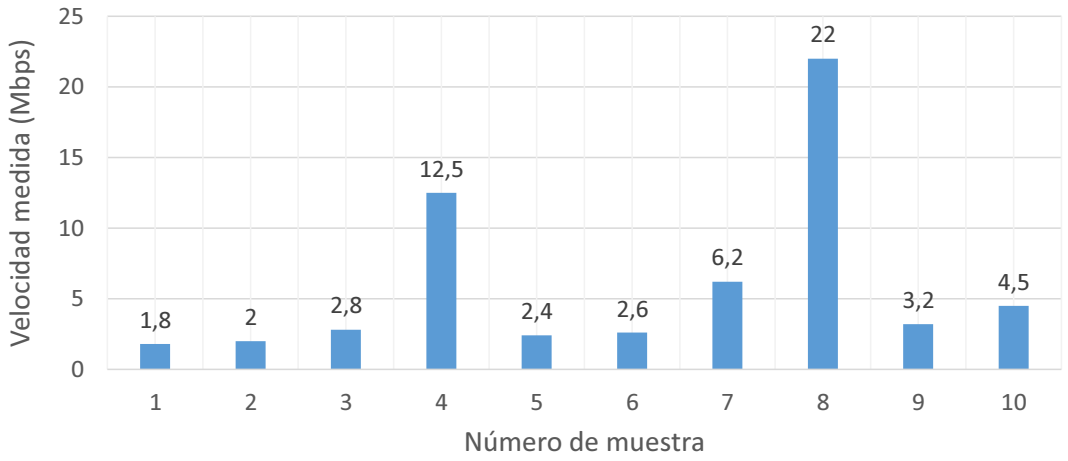

Figura 2.12: Ejemplo de medidas de velocidad en una red de acceso

tanto, la capacidad de una red como la habilidad de transportar tráfico de red de un conjunto de usuarios.

Los términos de velocidad y capacidad se encuentran relacionados. La figura 2.13 muestra la relación entre la velocidad y la capacidad de la red, ilustrando los diferentes tipos de velocidades que pueden definirse en función de ciertas consideraciones:

- Tráfico medio. Los usuarios envían o reciben tráfico de forma intermitente, por lo que el tráfico medio entre todos los suscriptores, incluyendo a los activos e inactivos, siempre es mucho menor que la velocidad actual a la que el tráfico es transferido a un usuario activo determinado.

- Capacidad promediada. Esta capacidad promediada por el número de usuarios, incluyendo activos e inactivos, puede verse como la parte que le correspondería a cada suscriptor del ancho de banda disponible en la red. En el caso en el que todos los suscriptores de la red estuviesen activos de forma simultánea, la velocidad instantánea percibida por los suscriptores seria igual a la capacidad promediada.

- Velocidad sostenible. Esta velocidad es la tasa de transferencia media entre los suscriptores que se encuentran activos. Si la red de comunicaciones se encuentra diseñada para soportar el tráfico medio de todos los suscriptores (activos e inactivos) con un margen para las ráfagas de tráfico que puedan existir, la velocidad sostenible suele ser mayor que la capacidad media por suscriptor.

- Velocidad actual. Se corresponde con el valor de velocidad instantánea experimentada por un usuario, que se corresponde con la definición de la sección anterior.

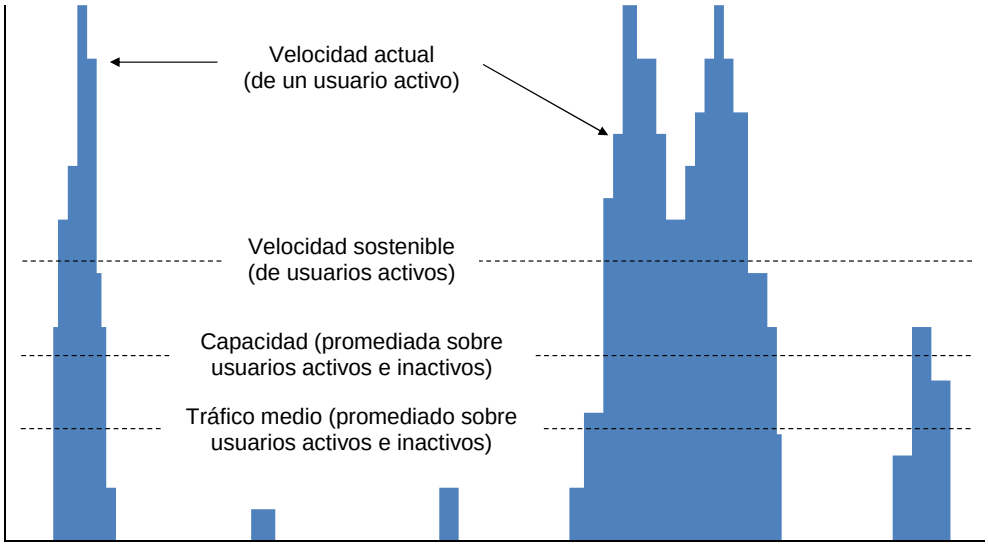

Figura 2.13: Relación entre velocidad y capacidad

Debido a que los suscriptores nunca demandan ancho de banda de forma simultánea, la velocidad actual de un usuario activo es generalmente mayor que la capacidad media.

Como también puede observarse en la figura, la velocidad de un usuario activo puede cambiar de un momento a otro ya que depende de los cambios de su propia demanda de tráfico y también del resto de suscriptores de la red.

2.3.1.3. Definición de capacidad de una red de acceso

La capacidad de una red de acceso se define de forma cuantitativa como el ancho de banda disponible existente entre el núcleo de la red (Internet) y los usuarios que están siendo servidos por la red de acceso. El ancho de banda disponible se encuentra limitado por el enlace de menor ancho de banda que exista entre los suscriptores y el núcleo de la red.

No obstante, definir de forma cuantitativa la capacidad de una red de acceso no siempre es sencillo, pues la red de acceso puede estar diseñada de forma que existan suscriptores con diferente capacidad. Por ejemplo, en la figura 2.14, se observa como en la red algunos suscriptores se encuentran conectados en niveles de agregación diferentes.

- EL nodo A tiene un enlace de 10 Gbps compartido por 200 suscriptores, 100 que están directamente conectados al nodo y otros 100 a través del Nodo B. La capacidad prorrateada entre los 200 suscriptores es de 50 Mbps por suscriptor.
- El nodo B tiene un enlace de 1 Gbps y 100 suscriptores conectados a través de

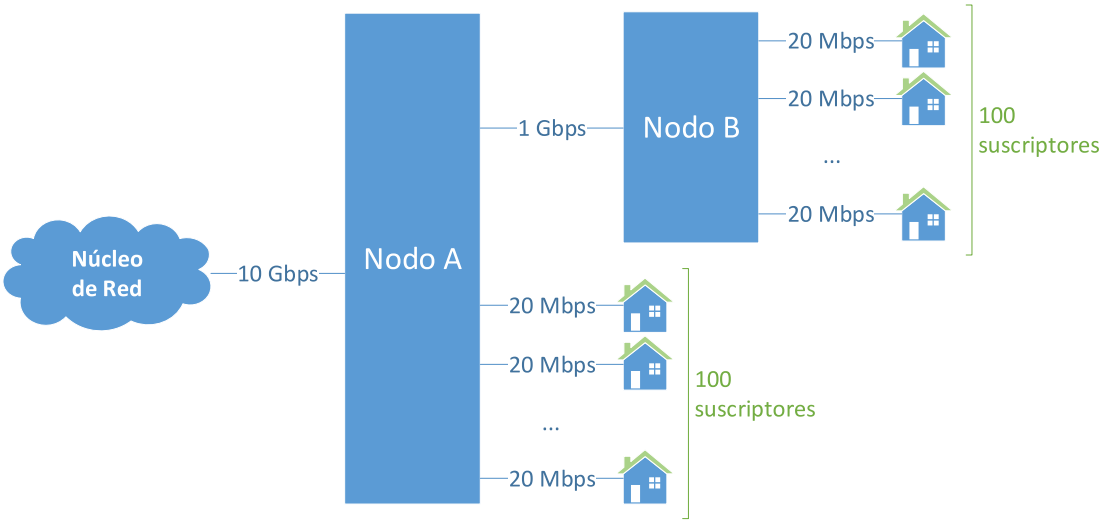

Figura 2.14: Ejemplo de capacidad de red

enlaces de 20 Mbps. La capacidad prorrateada es de 10 Mbps por suscriptor.

A partir de estos cálculos, se puede concluir que en la sección de red correspondiente al *Nodo A*, la limitación proviene de los propios enlaces de los suscriptores, siendo su capacidad de 20 Mbps. Por el contrario, los suscriptores de la sección de red del *Nodo B*, el enlace que está limitado es el correspondiente a 1 Gbps, por lo que la capacidad por suscriptor es de 10 Mbps.

2.3.1.4. Grado de Servicio (GoS)

Debido a que el rendimiento de la red no puede ser cuantificado mediante parámetros como la velocidad o la capacidad de la red, surge la necesidad de buscar algún tipo de métrica que relacione la velocidad sostenible o percibida que puede experimentar un suscriptor de una red a lo largo del tiempo.

En [Adtran, 2009] se hace manifiesta la necesidad de proveer algún tipo de métrica representativa del rendimiento de red desde el punto de vista del usuario. En este artículo se introduce el concepto de velocidad sostenible, es decir, la velocidad que puede experimentar un usuario con el 99 % de probabilidad. Así pues, se introduce la probabilidad de experimentar una velocidad a partir de una capacidad de la red de acceso.

Este concepto ya era conocido en el campo de la ingeniería de tráfico para las redes de voz basadas en conmutación de circuitos. En este contexto, el GoS se define como la probabilidad en la que una llamada en un grupo de circuitos es bloqueada o retrasada más de un umbral determinado. Además, el término de GoS ha estado muy relacionado

con la *hora cargada*, pues es cuando existe mayor intensidad de tráfico, también en las redes convencionales basadas en conmutación de circuitos. Así pues, el GoS de un sistema que tenga pérdidas, en este caso de llamadas, sigue la ecuación (2.23). Este parámetro oscila entre 0 % y 100 %, donde 0 % implica que no se perdería ninguna llamada y 100 % que se perderían todas.

$$GoS = \frac{\text{número de llamadas perdidas}}{\text{número de llamadas ofrecidas}} \tag{2.23}$$

En sistemas sin pérdidas, el GoS no puede definirse siguiendo la ecuación anterior y se utilizan tres medidas diferentes para caracterizar el rendimiento del sistema:

- El tiempo medio en el que un usuario espera una conexión si su llamada ha sido retrasada.

- El tiempo medio en el que un usuario espera una conexión sin importar si su llamada ha sido retrasada o no.

- La probabilidad que una llamada de un usuario pueda ser retrasada más de un tiempo umbral t mientras espera una conexión. Este umbral es seleccionado de forma que se pueda medir si el servicio cumple con un determinado GoS.

A partir de estos ejemplos, se puede comprender que la velocidad sostenible introducida en [Adtran, 2009] es similar al GoS descrito para sistemas sin pérdidas. En el caso de las redes de acceso, se busca la velocidad que puede ser obtenida por los usuarios con una probabilidad prefijada, como por ejemplo, el 99 % de los casos.

En definitiva, este enfoque es acertado a la hora de medir el rendimiento de una red de acceso, pues caracteriza el rendimiento obtenido por los usuarios para una probabilidad o un conjunto de probabilidades determinadas. No obstante, hay que tener en cuenta que, en una red de conmutación de paquetes hay factores que pueden afectar al cálculo del GoS, como por ejemplo, cómo se comparten los recursos de red compartidos entre los usuarios de la misma.

2.3.2. Arquitecturas de red de acceso

En esta sección se hace un análisis de la arquitectura presente en las redes de acceso más relevantes [Álvarez-Campana et al., 2009]. Se describen aquellas características de especial interés que afectan a la capacidad percibida por los suscriptores. En todas las arquitecturas descritas a continuación, el alcance del análisis de la capacidad se limita a la red de acceso entre el suscriptor y la puerta de enlace a Internet.

Familia	Rec. ITU-T	Fecha	Cap. Bajada	Cap. Subida
ADSL	G.992.1	1999	7 Mbps	800 Kbps
ADSL2	G.992.3	2002	8 Mbps	1 Mbps
ADSL2+	G.992.5	2003	24 Mbps	1 Mbps

Tabla 2.21: Tecnologías ADSL: recomendaciones, fecha y capacidades máximas

2.3.2.1. Línea de abonado digital (DSL)

A pesar de que este término puede referirse a varias tecnologías, en la mayoría de las veces hace referencia a Asymmetric Digital Subscriber Line (Línea de Abonado Digital Asimétrica) (ADSL), que es la tecnología instalada con mayor frecuencia. Por esta razón, a continuación se describe con mayor detalle ADSL, algunas de sus evoluciones y la última recomendación de DSL, Very high bitrate Digital Subscriber Line (Línea de Abonado Digital de Muy alta velocidad binaria) (VDSL).

ADSL. ADSL es la tecnología de banda ancha más utilizada actualmente en sus distintas variantes más relevantes: ADSL, ADSL2 y ADSL2+. Desde los primeros despliegues, los sistemas ADSL han sufrido una evolución muy importante en la tecnología de transporte físico (de ADSL a ADSL2+), en su arquitectura de red de acceso (de Asynchronous Transfer Mode (Modo de Transferencia Asíncrona) (ATM) a Ethernet) y en su aplicación (del acceso a Internet a moderada velocidad al soporte de ofertas triple play con distribución de TV, vídeo bajo demanda y un aumento importante en la velocidad del acceso a Internet).

En la siguiente tabla 2.21 se presentan las variantes de mayor relevancia de ADSL junto con las recomendaciones ITU Telecommunication Standardization Sector (Sector de Normalización de las Telecomunicaciones de la ITU) (ITU-T) donde se especifican, la fecha de edición de las mismas y sus capacidades máximas [ITU-T, 1999, ITU-T, 2002, ITU-T, 2003a].

VDSL se trata de una evolución de la tecnología ADSL con la principal característica de dotar al suscriptor con una tasa de transferencia muy alta. Para bucles de longitud reducida, es posible extender los límites de la tecnología ADSL, empleando frecuencias de hasta 30 MHz sobre el par de cobre. Para conseguir bucles de menor longitud, la tecnología VDSL va acompañada de un despliegue de fibra hasta los nodos de acceso, desde los cuales se alcanza al abonado, de forma que así se reduzca la distancia del par de cobre. En la tabla 2.22 se muestran las dos recomendaciones del ITU-T junto con sus capacidades de bajada y subida [ITU-T, 2001, ITU-T, 2006].

Arquitectura de red. En una red de acceso DSL, la arquitectura de red suele ser similar a pesar de las diferentes tecnologías que existen. No obstante, tal y como se

Familia	Rec. ITU-T	Fecha	Cap. Bajada	Cap. Subida
VDSL	G.993.1	2001	55 Mbps	3 Mbps
VDSL2	G.993.2	2006	100 Mbps	100 Mbps

Tabla 2.22: Tecnologías VDSL: recomendaciones, fecha y capacidades máximas

aprecia en la figura 2.15, puede darse el caso de que la arquitectura consista de 2 o 3 fases diferentes:

1. La red de media milla entre la puerta de enlace a Internet (Internet Gateway) y la Central Office (Oficina Central) (CO), se compone, habitualmente, de nodos de agregación de considerable tamaño y multi-servicio. En el caso de DSL, estos nodos suelen ser DSL Access Multiplexers (Multiplexores de Acceso DSL) (DSLAMs) modulares compuestos a su vez por tarjetas de acceso. Estos DSLAMs de CO suelen dar acceso y soporte a miles de suscriptores DSL, bien de forma directa o a través de conexiones de segunda milla a través de otros DSLAMs. La conexión entre Internet Gateway y CO suele ser a través de uno o varios enlaces de varios Gbps o, incluso de mayor capacidad.

2. La red de segunda milla se encuentra entre COs y DSLAMs remotos en planta exterior. Estos DSLAMs suelen estar montados en locales remotos o cabinas de intemperie. Su número y localización depende de la tecnología DSL utilizada, ya que éstas pueden requerir unas distancias del bucle de par de cobre determinada. La longitud de estos bucles entre abonados y DSLAMs se reducen cuando estos últimos se colocan en la planta exterior, lo cual permite que puedan haber tasas de transferencia más altas en la red de última milla. A pesar de que existen infinidad de soluciones de fabricantes, un DSLAMs suele dar servicio de acceso de 24 a 384 suscriptores. Antiguamente, los enlaces de segunda milla podían ser de cobre o de fibra, aunque actualmente, la tendencia es la de disponer de enlaces basados en fibra óptica.

3. La última milla se compone del bucle de abonado de par de cobre trenzado. Este bucle de abonado ofrece a cada suscriptor de una conexión dedicada hasta el DSLAM. Como se ha mencionado con anterioridad, la última milla se encuentra localizada en planta exterior (en caso de que existan DSLAMs desplegados) o directamente en la CO (en caso de que la distancia del bucle de abonado hasta el CO sea suficientemente corta).

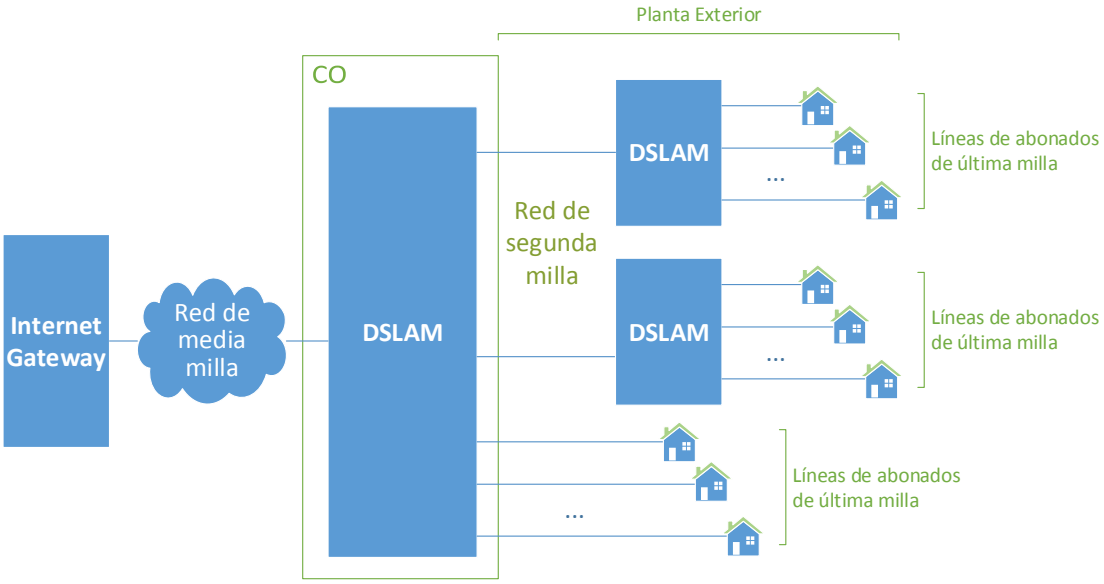

Figura 2.15: Arquitectura de red de acceso DSL

2.3.2.2. Hybrid Fiber Coaxial (HFC)

Las redes de acceso basadas en Hybrid Fiber Coaxial (Híbrido de Fibra-Coaxial) (HFC) se basan en la combinación de redes de fibra óptica y redes de cable coaxial. Estas redes son el resultado de la evolución de las redes de distribución de televisión por cable, ya que permiten el acceso a Internet de banda ancha utilizando las redes Community Antenna Television (Televisión por cable) (CATV). Una característica representativa de esta tecnología es que la última milla, basada en cable coaxial, es un recurso de red compartido por todos los abonados de una zona determinada.

En la figura 2.16 se muestra la figura de la arquitectura de una red HFC bidireccional. Una red HFC típica se compone de los siguientes elementos principales: cabecera, red troncal, red de distribución y los equipos de abonado.

La función principal de la cabecera HFC es combinar distintas fuentes de programación de televisión y ubicarlas en el espectro del cable. Por ejemplo, la televisión digital se encuentra ubicada en canales en la banda entre 55 y 550 Mhz. El transporte de datos se ubica la banda de 5 a 50 MHz en sentido ascendente y en la banda de 550 a 860 MHz en sentido descendente. El elemento que recibe y envía los flujos de datos en la cabecera es el Cable Modem Termination System (Sistema de Terminación de Cablemódems) (CMTS), el cual se encarga de codificar, modular y gestionar el acceso al medio compartido por los módems de cable.

Las redes troncales HFC son redes ópticas con topologías en anillo de varios niveles (dos o tres niveles normalmente) que se encuentran entre las cabeceras (CMTS) y los

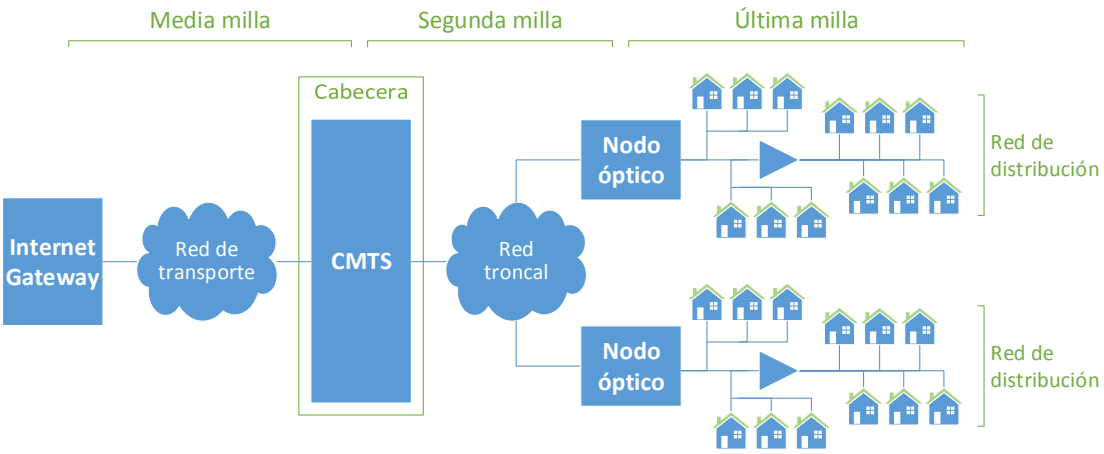

Figura 2.16: Arquitectura de red de acceso HFC

nodos ópticos. Los nodos ópticos son responsables de realizar la conversión entre señales ópticas y eléctricas para el tráfico de bajada, y la conversión inversa para el tráfico de subida. Estos nodos suelen estar localizados cercas de los abonados.

La red de distribución HFC es una red de transporte basada en cable coaxial cuyos elementos son:

- **Amplificador de línea:** amplificadores con requisitos muy estrictos en cuanto al ancho de banda a amplificar y a la potencia suministrada (especialmente en la parte del espectro del sentido descendente). Estos amplificadores se suelen tele-alimentar desde el nodo óptico.
- **Terminal Access Point (Punto de Acceso de Terminal) (TAP):** estos elementos se encargan de derivar parte de la energía que se transmite por el coaxial hacia las terminaciones donde se conectan las acometidas de los usuarios.
- **Divisores de potencia:** permiten derivar ramales coaxiales.

El equipamiento que un abonado de red HFC necesita para acceder a los servicios que ésta provee, está formado por un módem de cable y un *set-top box*. Los módems de cable actúan de pasarela entre la red del cliente (Ethernet normalmente) y la red de cable. Los módems del usuario, junto a los de la cabecera, implementan los niveles físico y de control de acceso al media (MAC) propias para la tecnología HFC. Los *set-top boxes* son dispositivos que se conectan entre la toma coaxial y el receptor de TV. Estos dispositivos disponen de una etapa de demodulación específica del medio de transporte, efectuando también el procesado de la codificación e información de sistema MPEG/DVB y el descifrado en el caso de canales de pago.

Estándar	Prestaciones	Vel. Bajada	Vel. Subida
DOCSIS 1.X	Primera especificación Best effort	43 Mbps	10 Mbps
DOCSIS 2.0	Calidad de Servicio Bajo coste	43 Mbps	30 Mbps
DOCSIS 3.0	Channel bonding Soporte de IPv6	$m \cdot 43$ Mbps	$n \cdot 30$ Mbps

Tabla 2.23: Estándar DOCSIS: características principales

Data Over Cable Service Interfaces Specification (Especificación de Interfaz para Servicios de Datos por Cable) (DOCSIS). El estándar más importante en el mundo de las redes de cable es DOCSIS, desarrollado por *CableLabs* y aceptado como estándar por organismos como ITU, European Telecommunications Standards Institute (Instituto Europeo de Normas de Telecomunicaciones) (ETSI) y Society of Cable Telecommunications Engineers (Sociedad de Ingenieros de Telecomunicaciones de Cable) (SCTE). Existen varias versiones de DOCSIS cuyas principales características (prestaciones y velocidades máximas) se citan en la tabla 2.23 y se corresponden con las especificaciones descritas en [CableLabs, 1996].

ETSI publica el estándar [ETSI, 2003a] (conocido como *euroDOCSIS*) en el que se particularizan algunos aspectos de DOCSIS al contexto europeo. Por otro lado, el estándar [ETSI, 2003b] (conocido como *IPCablecom*) especifica un conjunto de protocolos y requisitos funcionales asociados, desarrollados para proporcionar servicios IP multimedia seguros con requisitos exigentes de QoS sobre redes HFC.

En cuanto a las prestaciones de HFC, el caudal de datos en DOCSIS 2.0 puede ofrecer tasas de hasta 38 Mbps para el canal de bajada y hasta 27 Mbps por cada canal de 6 Mhz de subida. En un despliegue residencial típico se suele reservar uno o dos canales para la transmisión de datos en sentido descendente. Debido a limitaciones propias del medio físico de esta tecnología, la capacidad de subida en sistemas actuales suele ser del orden de 35 Mbps para servir cerca de 250 abonados [Limaye et al., 2008]. Una de las principales ventajas que aporta DOCSIS 3.0 es que permite agrupar varios canales, multiplicando las tasas de bit que se pueden conseguir. Además de los límites impuestos por los canales compartidos, la tasa de transferencia también puede verse limitada debido al equipamiento de usuario, es decir, por el módem de cable. Es más probable que esta limitación suceda en sentido de bajada que en el de subida.

2.3.2.3. Fiber To The Home (FTTH)

El acceso basado en fibra, es la tecnología que mejores prestaciones ofrece desde el punto de vista técnico. Entre las muchas virtudes de la FTTH, destaca su gran

Estándar	Rec. ITU-T	Vel. Bajada	Vel. Subida	*Splitting*	Alcance
GPON	G.984.1	2,5 Gbps	1,25 Gbps	64/128	20 km
XGPON	G.987.1	10 Gbps	2,5 Gbps	64/128/256	20 km

Tabla 2.24: Recomendaciones más relevantes FTTH: características principales

capacidad en términos de ancho de banda, sus bajas pérdidas de transmisión, su facilidad de despliegue en planta y su resistencia a las agresiones del entorno. A pesar de que existen multitud de configuraciones de acceso de fibra dependiendo de la cercanía de la fibra al usuario, esta tesis se centra en el acceso de fibra hasta el mismo domicilio del abonado (FTTH).

Passive Optical Network (Red Óptica Pasiva) (PON). Las redes ópticas pasivas permiten la utilización de una única fibra, para dar servicio a un conjunto de usuarios, sin utilizar regeneración óptica y consiguiendo buenos rangos de alcance. A pesar de que existen varias tecnologías disponibles para el acceso basado redes ópticas pasivas, esta tesis doctoral se centra en la tecnología Gigabit-capable Passive Optical Network (Red Óptica Pasiva con Capacidad de Gigabit) (GPON), definida en la recomendación ITU-T G.984 [ITU-T, 2003b]. El estándar GPON permite la explotación de las redes PON hasta regímenes de 2488 Mbps, soportando protocolos Ethernet, ATM y Time Division Multiplexing (Multiplexación por División de Tiempo) (TDM). En 2010, la recomendación de la ITU-T G.987 [ITU-T, 2010] define el nuevo estándar XG-PON o 10G-PON que alcanza tasas de bit de hasta 10 Gbps. Estas dos tecnologías se encuentran descritas en la tabla 2.24, donde se muestran también las capacidades que ofrecen, las divisiones y su alcance.

Los principales elementos de los que consta una GPON son:

- **Optical Line Termination (OLT):** Terminador óptico de línea, ubicado en la cabecera de red, es decir, en la central del operador. Cada OLT puede dar servicio a varias PONs. El nivel de *splitting* no puede ser muy elevado, siendo frecuente la utilización de uno o dos niveles de 1:8 (consiguiendo hasta 64 usuarios) o un nivel 1:32 (consiguiendo alcanzar 32 usuarios).
- **Optical Network Termination (ONT):** Terminadores ópticos de red, ubicados en las instalaciones del abonado.
- **Optical Distribution Network (ODN):** Red de distribución óptica, formada por cables de fibra, divisores pasivos y acopladores.

En la figura 2.17 se muestra una red típica GPON, sus componentes principales y las partes en las que se puede descomponer la arquitectura de red FTTH:

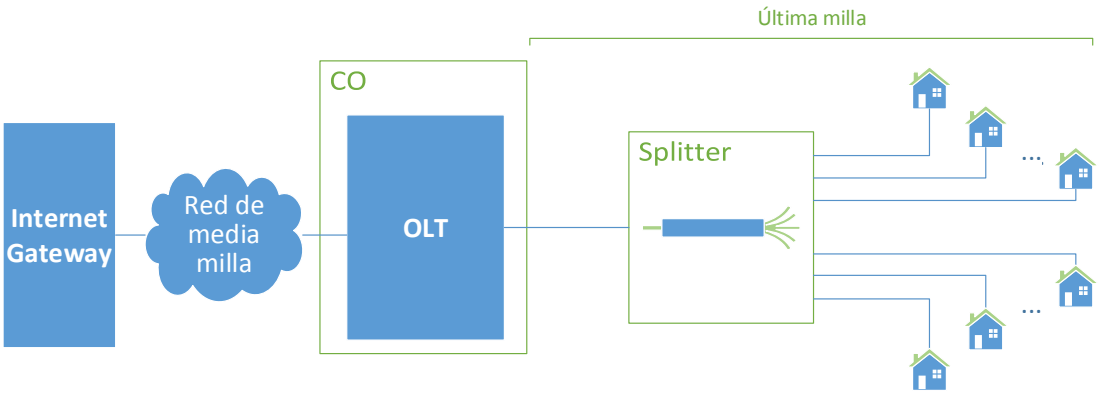

Figura 2.17: Arquitectura de red de acceso FTTH GPON

1. La red de media milla se encuentra entre la puerta de enlace a Internet y el OLT. Al igual que en otras arquitecturas de red, los enlaces en esta red suelen ser de alta capacidad con anchos de banda de varios Gbps.

2. La última milla se encuentra entre el OLT y los ONTs que se encuentran localizados en los domicilios de los abonados o muy cercanos. Los divisores de fibra (*splitters*) separan de forma óptica las señales de bajada y unen las señales de subida de las fibras individuales de los abonados.

Las comunicaciones de datos de subida y de bajada de una GPON se realizan en diferentes longitudes de onda. Como se ha descrito con anterioridad, la capacidad de bajada es de 2,5 Gbps, la cual ha de ser compartida entre todas las ONTs. El enlace de subida es compartido por todos los suscriptores mediante el uso de Time Division Multiple Access (Acceso Múltiple por División de Tiempo) (TDMA) y tiene una capacidad de 1,25 Gbps. Existen algunos despliegues en los que se incluye una tercera longitud de onda para la provisión de servicios de video.

La tasa máxima disponible para cada abonado también depende del equipamiento de usuario. Por ejemplo, si un usuario se conecta a la ONT mediante una interfaz inalámbrica, su tasa máxima puede estar limitada por la tecnología inalámbrica utilizada.

2.3.2.4. Broadband Wireless Access (BWA)

El Broadband Wireless Access (Acceso Inalámbric de Banda ancha) (BWA) es otro ejemplo donde la red de última milla también es un recurso compartido entre varios usuarios. Los suscriptores de la última milla comparten un mismo canal utilizando mecanismos y protocolos de acceso múltiple. A pesar de que existen numerosos despliegues de redes BWA basadas en tecnologías Wi-Fi u otras soluciones propietarias,

las redes de tipo BWA más relevantes se corresponden a las tecnologías basadas en LTE o Worldwide Interoperability for Microwave Access (Interoperabilidad Mundial para Acceso por Microondas) (WiMAX). Estas dos tecnologías, a pesar de haber sido concebidas para acceso inalámbrico con soporte a la movilidad, también pueden dar servicio a abonados de banda ancha fijos.

En la figura 2.18 se muestra una visión general de una red BWA que podría estar basada en cualquier tecnología inalámbrica, como por ejemplo las mencionadas LTE o WiMAX. Esta red está compuesta por 3 secciones entre la puerta de enlace a Internet y los suscriptores:

1. La red de media milla se encuentra entre la puerta de enlace a Internet y los nodos de agregación. Estos nodos se encuentran etiquetados en la figura como *Pasarelas de Servicio* y su denominación puede cambiar en función de la tecnología utilizada para el acceso al medio, como por ejemplo, *Controladores de Red de Radio* o *Controladores de Estaciones Base*.

2. La red de segunda milla se sitúa entre los nodos de agregación anteriores y las estaciones base celulares, las cuales incluyen equipamiento de transmisión inalámbrica, amplificadores y antenas. Los enlaces de estas redes pueden de diversa naturaleza, pues podrían ser inalámbricas (enlaces punto-a-punto utilizando WiMAX u otra tecnología propietaria) o fijas (fibra óptica, DSL, etc.).

3. Las redes de última milla son aquellas que se encuentran entre las estaciones base y los suscriptores. Cada estación base da una determinada cobertura a un conjunto de usuarios. A esta zona geográfica de cobertura también se le conoce como celda. Una red inalámbrica puede estar compuesta por numerosas celdas y sus suscriptores se comunican con aquellas estaciones base que le proporcionen mejor conectividad inalámbrica.

A continuación, se introducen las principales características de redes BWA basadas en las tecnologías inalámbricas LTE y WiMAX.

Long Term Evolution (LTE). LTE es un estándar que define la arquitectura de red y el acceso de radio para comunicaciones de alta velocidad dispositivos móviles y otros terminales. LTE ha sido desarrollado por el 3rd Generation Partnership Project (3GPP) y se encuentra especificado en su *Release 8* [ETSI, 2009] con algunas mejoras en la *Release 9* [ETSI, 2011b]. En el estándar se utilizan canales con anchos de banda flexibles entre las frecuencias de 1,4 MHz hasta 20 MHz.

Las principales novedades en esta tecnología es que la interfaz radioeléctrica se encuentra basada en Orthogonal Frequency-Division Multiple Access (Multiplexación por División de Frecuencia Ortogonal) (OFDMA) en el sentido descendente y en Single

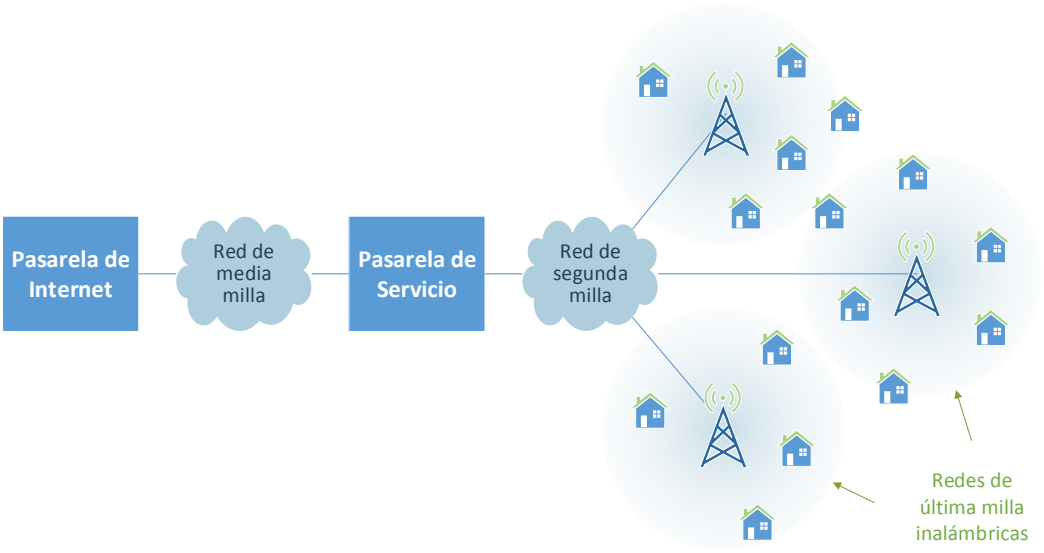

Figura 2.18: Arquitectura de red de acceso BWA

Carrier Frequency Divison Multiple Access (Acceso Múltiple por División de Frecuencia de Portadora Única) (SC-FDMA) en el ascendente. Ambas tecnologías permite que los datos de distintos suscriptores pueda ser multiplexada en el dominio del tiempo y de la frecuencia. Las transmisiones de datos en el plano de usuario tienen una latencia muy baja debido a que se encuentran agrupadas en frames de radio de 10 ms, los cuales también se encuentran sub-divididos en subframes de 1 ms. A pesar de que las velocidades dependen de la categoría de equipamiento de usuario, las tasas de bit pueden llegar a valores cercanos a 300 Mbps en sentido descendente y 75 Mbps en sentido ascendente.

LTE-Advanced se estandariza en la *Release 10* del 3GPP [ETSI, 2011a] y propone una serie de mejoras al anterior estándar de comunicaciones móviles LTE. Una de las principales características de este nuevo estándar es la capacidad de tomar ventaja de la topología avanzada de las redes, para así, poder optimizar las comunicaciones mediante el uso de nuevas estructuras de red, como por ejemplo, picoceldas, femtoceldas y nuevos nodos de transmisión. LTE-Advanced hace un uso más eficiente del espectro y supera las prestaciones de su predecesor con tasas de bit máximas de 1 Gbps en sentido descendente y de 500 Mbps en sentido ascendente.

WiMAX. WiMAX es una tecnología de transmisión de datos a través de ondas de radio en las frecuencias de 2,4 GHz y que puede tener una cobertura de hasta 50 km. WiMAX es una tecnología de bucle local utilizada para las redes de última milla, que se

encuentra definida bajo el estándar del IEEE 802.16 [IEEE, 2002]. La ventaja de este estándar reside en su capacidad de proporcionar acceso a Internet de alta velocidad sin que los suscriptores tengan que estar en línea de visibilidad con las estaciones base.

Actualmente existen dos variantes del estándar:

- **WiMAX Fijo:** Se encuentra definido en [IEEE, 2004] y se basa en enlaces radio entre estaciones base y equipos de usuario situados en los domicilio de los suscriptores.
- **WiMAX Móvil:** Se encuentra definido en [IEEE, 2006] y permite el desplazamiento de los usuarios de forma similar a la que sucede en redes de telefonía móvil.

Las velocidades teóricas máximas de WiMAX dependen del ancho de banda del canal y de la modulación utilizada, oscilando entre los 32 Mbps y los 134 Mbps. La tecnología de WiMAX hace uso de OFDMA en frames de radio de 5 ms, tanto para el sentido descendente como ascendente. Este tipo de acceso al medio permite permite que los datos de subida y bajada de los diferentes usuarios de puedan ser multiplexados tanto en el dominio del tiempo como de frecuencia.

Características de BWA. Tanto en las redes de tipo LTE como WiMAX incluyen el soporte para algunas técnicas Multiple-Input Multiple-Output (Múltiple Entrada Múltiple Salida) (MIMO) que pueden ser utilizadas para aumentar las prestaciones de las comunicaciones inalámbricas, como por ejemplo, el aumento del rendimiento de transferencia de datos a algunos usuarios específicos, soporte de usuarios adicionales o incluso el aumento en la cobertura inalámbrica. Estas técnicas también pueden ser combinadas con otras, como por ejemplo la asignación dinámica de sub-canales para permitir un factor de reutilización de frecuencia de 1, es decir, compartir la totalidad del espectro de frecuencias de forma dinámica en todas las celdas de la red.

Habitualmente, a los máximas teóricos para las tecnologías de LTE y WiMAX hay que restarle el *overhead* correspondiente a los niveles físicos y de enlace. Además, en una red inalámbrica la velocidad percibida por un suscriptor siempre es menor a la máxima teórica, ya que el medio es compartido por los otros usuarios de la celda y porque sólo aquellos usuarios muy próximos a las estaciones base podrán disfrutar de tasas de bit máximas. Las tasas de bit ofrecidas por una red inalámbrica aumentan con el cuadrado de la distancia hasta la estación base.

Existen muchos otros parámetros que pueden limitar el rendimiento de una red inalámbrica, entre los que destacan los siguientes:

- Atenuación de señal: ocasionada principalmente por la distancia a la estación base, aunque también le afectan otros factores como obstáculos, reflexiones de señales,

etc.

- Interferencias: pueden ser resultado de la existencia de otras celdas que utilicen los mismos canales de frecuencias. Tanto LTE como WiMAX utilizan técnicas de procesado de señal complejas en el dominio de la frecuencia que permiten utilizar los mismos canales en celdas adyacentes, mejorando de esta forma la eficiencia espectral de la red.

- Número y tipo de suscriptores: debido a que las transmisión de datos se realiza a través de un medio compartido, el número y tipo de usuarios afecta al rendimiento de la red.

- Equipamiento de usuario: en algunos casos particulares puede darse el caso de que el equipamiento de usuario limite el rendimiento de la red en términos de tasa de datos máxima que pueden soportar.

2.3.3. Modelo genérico de arquitectura de red de acceso

Las redes de acceso analizadas anteriormente pueden verse como una serie de enlaces que llegan hasta las instalaciones del usuario y cuyo tráfico hacia Internet, se va combinando en una serie de niveles de agregación. En el sentido contrario, el tráfico de bajada de los usuarios se va repartiendo en cada nivel de agregación hasta llegar al enlace del usuario destino. Partiendo de esta idea, un posible esquema de referencia sería el que se muestra en la figura 2.19. En el tramo desde la pasarela de Internet hasta el equipo de usuario del abonado pueden existir un número indeterminado de niveles de agregación, caracterizado a partir de un ratio de agregación de *1:x*. Para el caso de las arquitecturas BWA, el domicilio del abonado representado en la figura podría ser directamente un dispositivo móvil, como por ejemplo, un teléfono móvil.

A partir del análisis en detalle de cada una de las arquitecturas de redes de acceso descritas, se puede concluir que éstas comparten características de agregación en los diferentes tramos de la red. La figura 2.20 muestra el esquema de referencia orientativo en el que se desglosa la red de acceso en diferentes tramos y niveles de agregación. A pesar de que cada arquitectura de red puede tener sus propias particularidades, estas características y factores de agregación suelen venir determinados por las infraestructuras (edificios) disponibles y los proveedores de Internet, siendo muy similares entre diferentes tecnologías de acceso.

Debido a que cada tecnología de acceso tiene su propia terminología y ciertas peculiaridades en algunos tramos de agregación, se puede realizar una abstracción en 3 etapas de forma que el nivel de agregación que en cada una de estas etapas sea similar. El resultado de esta identificación es el modelo conceptual representado en la figura 2.21 y que cuenta con las siguientes etapas:

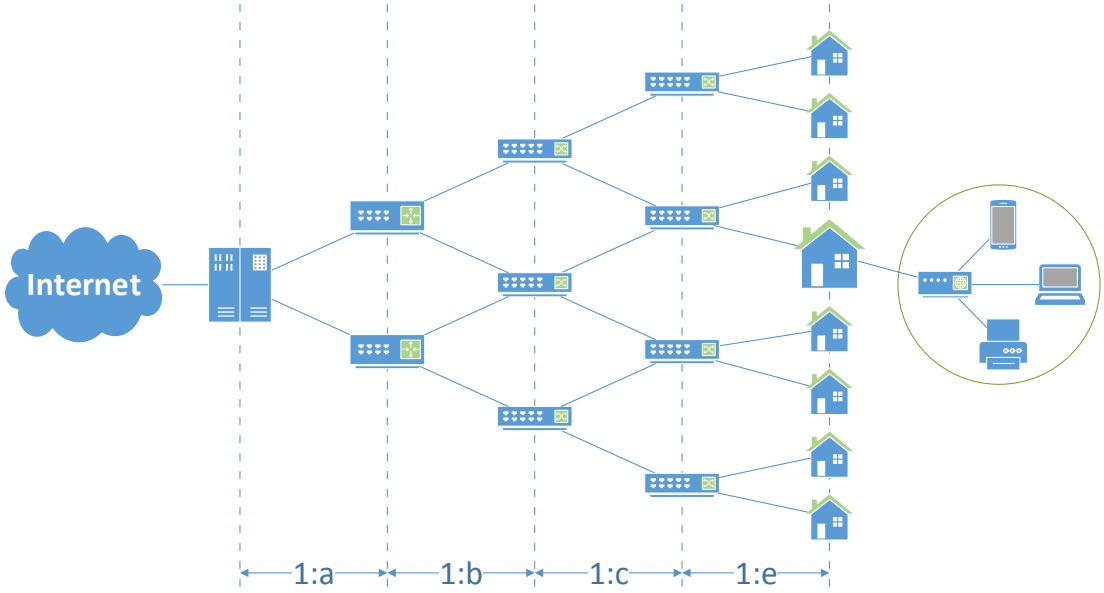

Figura 2.19: Modelo genérico de arquitectura de red de acceso

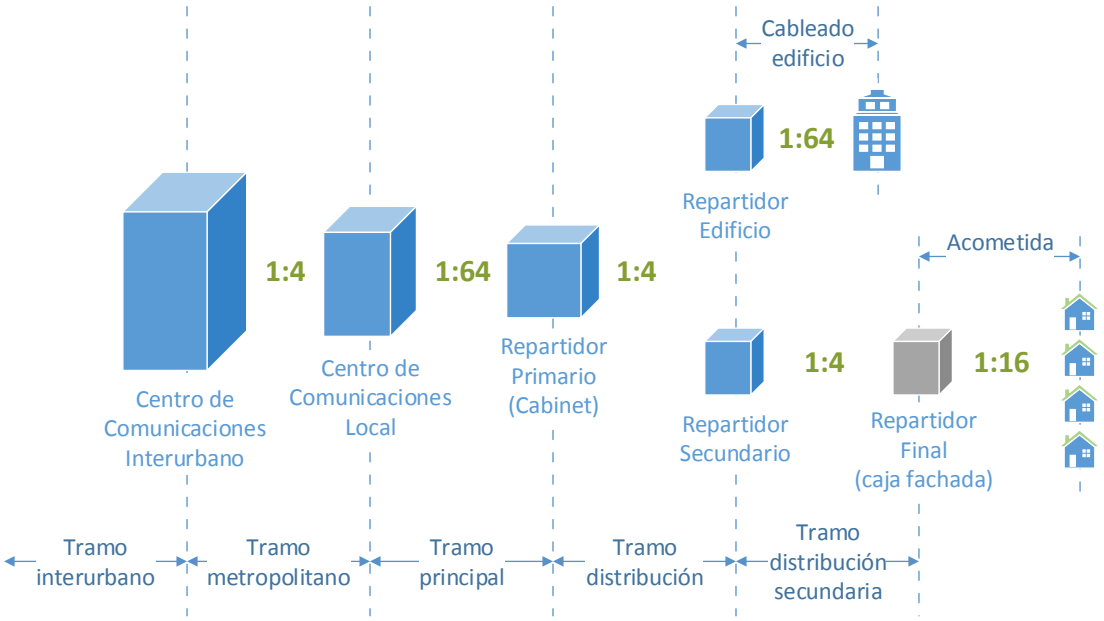

Figura 2.20: Esquema de referencia de redes de acceso

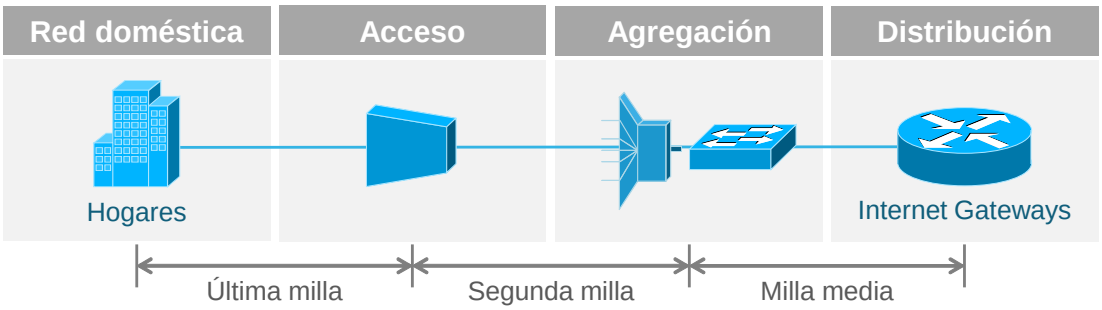

Figura 2.21: Modelo genérico de arquitectura de red de acceso basado en 3 etapas

1. **Media milla.** Esta etapa se encuentra entre la puerta de enlace a Internet (Internet Gateway) y los nodos de agregación, los cuales agregan el tráfico de una gran cantidad de suscriptores en enlaces compartidos con un gran ancho de banda. Estos nodos de agregación se encuentran desplegados en un número pequeño de localizaciones, como por ejemplo, en las COs de un operador. Este tipo de nodos de agregación y sus interfaces de comunicaciones pueden ser específicos del tipo de red de acceso. La media milla suelen ser redes de alta velocidad con velocidades que oscilan entre DS3 (45 Mbps) para nodos de agregación pequeños hasta múltiples gigabits por segundos para nodos más modernos.

2. **Segunda milla.** Esta red es la que se encuentra entre los nodos de agregación anteriores y los nodos de acceso. Los nodos de acceso son aquellos nodos más cercanos a los suscriptores. Estos nodos de acceso son totalmente específicos a la tecnología de acceso. Este tipo de nodos, realizan también una función de agregación de suscriptores. Este tipo de nodos incluyen DSLAMs, nodos ópticos o estaciones base inalámbricas. La localización de estos nodos puede ser variada, ya que pueden encontrarse en el exterior o incluso haber sido implementados en la misma pieza de equipamiento que los nodos de agregación. La capacidad de la segunda milla suele oscilar entre megabits hasta gigabits por segundo, dependiendo del tipo de red y del número de suscriptores servidos.

3. **Última milla.** Esta etapa corresponde a la red que se encuentra entre los suscriptores y los nodos de acceso más cercanos a los mismos. Son los primeros nodos de la red de acceso desde el punto de vista de los suscriptores. La naturaleza de esta red es totalmente dependiente de la tecnología de acceso utilizada. La capacidad de la última milla suele encontrarse entre varios megabits hasta los 100 Mbps para los despliegues de red más modernos.

Este modelo de referencia es utilizado para la caracterización de redes de acceso,

la cual es necesaria para aplicación de la metodología propuesta en esta tesis doctoral para la estimación de demanda en las redes de acceso (capítulo 4).

2.4. Conclusiones

En este capítulo se ha realizado una revisión del estado del arte de tres áreas de conocimiento bien diferenciadas, extrayendo para cada una de las mismas un conjunto de conclusiones.

El análisis del estado del arte sobre la caracterización de usuarios de Internet pone de manifiesto las siguientes conclusiones:

- Escasez de estudios en la literatura que hayan sido realizados siguiendo una metodología rigurosa para la extracción de tipologías de usuarios de Internet.
- Inexistencia de estudios recientes sobre los usuarios de Internet en España que extraigan un conjunto de perfiles en base a patrones de consumo de servicios.

Con objetivo de proveer una metodología rigurosa, se propone el uso de una metodología basada en KDPs para el descubrimiento y extracción de conocimiento a partir de fuentes de información. Además, con objeto de detallar algunas de las fases de la metodología, se realiza un análisis de las principales técnicas y métricas de calidad de minería de datos disponibles en la literatura.

El análisis exhaustivo sobre los modelos teóricos y estudios más relevantes de la literatura posibilitan una mejor comprensión del dominio del problema ligado a la extracción de perfiles de usuarios. Debido a que no existen estudios recientes en España, se describen las diferentes fuentes de información estadísticas disponibles para ser utilizadas por la metodología basada en KDP en el capítulo 3 de esta tesis doctoral.

Las principales conclusiones el análisis relativo a la caracterización de tráfico de Internet son las siguientes:

- Se ha realizado un análisis de diferentes modelos de tráfico con el objetivo de seleccionar un modelo para ser utilizado para estimar la demanda de tráfico. El modelo de fuente de tráfico seleccionado es de tipo ON/OFF *heavy-tail*.
- Se ha identificado una mezcla de tráfico de aplicaciones de Internet representativa de la mayor parte de tráfico de Internet. Las aplicaciones consideradas son: navegación web, compartición de ficheros, video sobre Internet y juegos en red.

A partir de la identificación de las aplicaciones de Internet representativas del tráfico de Internet, se ha realizado un análisis exhaustivo sobre los modelos de tráfico de aplicaciones de Internet disponibles en la literatura. A partir de este análisis, se propone un modelo de fuente de tráfico para cada aplicación de Internet considerada. Estos

modelos de tráfico serán utilizados posteriormente en la aplicación a casos de estudio, descritos en el capítulo 5 de esta tesis doctoral.

Por último, las conclusiones extraídas del análisis del estado del arte sobre el dimensionado de redes de acceso, son las siguientes:

- Se han descrito las diferentes métricas relativas al ancho de banda de una red, poniéndose de manifiesto la necesidad de definir una métrica que considere la probabilidad de obtener cierto rendimiento.
- A partir de la revisión de las principales arquitecturas de red de acceso más relevantes en la actualidad, se propone un modelo genérico de arquitectura de acceso independiente de la tecnología de red de acceso.

En el capítulo 4 se define una métrica de calidad de rendimiento que se basa en las conclusiones extraídas en esta sección del estado del arte. El modelo genérico de arquitectura de red es utilizado a lo largo de los capítulos 4 y 5 indistintamente.

Capítulo 3

Caracterización de usuarios de Internet

3.1. Introducción

Hoy en día, el usuario de Internet se encuentra sometido a una rigurosa vigilancia por parte de los proveedores de contenidos, especialmente en el caso de los usuarios de aplicaciones web. Uno de los objetivos más habituales consisten en realizar lo que se conoce como *ad targeting*, como por ejemplo para mostrar anuncios acordes con los intereses del usuario para maximizar así las probabilidades de éxito de la publicidad. Sin embargo, cuando se trata de medir el valor comercial de los usuarios, las grandes empresas depositan su confianza en estudios estadísticos tradicionales basados en paneles que representan a la población.

Las audiencias no pueden interpretarse como entidades que hacen uso de los servicios de una forma lineal. El consumo de éstos depende profundamente de las personas, de sus familias y de sus vidas de una forma realmente íntima y personal. Las caracterizaciones estadísticas proveen de una perspectiva amplia del universo que se quiere observar, lejos de la perspectiva sesgada que puedan disponer algunos análisis centrados en un sitio o plataforma.

En el caso particular de los usuarios de Internet, muchas veces se comete el error de pensar que la gran mayoría de usuarios se comportan de forma similar en cuanto al uso y consumo que realizan. No obstante, desde el punto de vista de los usuarios, Internet puede tener significados y propósitos muy diferentes, dando a lugar a distintos comportamientos de usuarios [Selwyn et al., 2005].

En la sección 2.1 se presentan los principales modelos teóricos de la literatura para comprender y explicar el consumo de las TICs, y en especial de Internet. Además, también se hace un riguroso análisis de los principales estudios relativos a la caracterización de usuarios de Internet, los cuales contribuyen a una mejor comprensión al dominio del

problema.

El objetivo principal de este capítulo es realizar una extracción de perfiles de usuarios de Internet mediante la identificación y clasificación de comportamientos y patrones de uso de servicios de Internet. Esta categorización de usuarios, también denominada tipología, no sólo refleja cómo son de diferentes los grupos de usuarios de Internet entre sí, sino que también describe cómo es el potencial de los tipos de usuario en referencia a sus preferencias de consumo de servicios.

En el contexto de esta tesis doctoral, este conocimiento extraído en forma de caracterización de usuarios de Internet es utilizado como parte de la metodología de estimación de demanda de tráfico en Internet. Su principal uso es la de caracterizar a los usuarios no sólo en sus hábitos de consumo de servicios, sino en el tráfico que generan, de forma que pueda ser utilizado para el dimensionamiento de algunos tramos de las redes de acceso.

3.1.1. Metodología empleada

Generalmente en los KDPs se parte de unos datos ya proporcionados [Han et al., 2012], los cuales se comportan como parámetros de entrada a la hora de aplicar conceptos y técnicas de minería de datos para descubrir información valiosa. Sin embargo, en el caso de esta tesis doctoral también se ha de realizar un análisis de diferentes fuentes de información con el objetivo de seleccionar aquella con mayor valor y rigurosidad a la hora de proporcionar los datos de entrada al KDP.

En la sección 2.1, se presentan varios modelos de KDP, los cuales comparten gran número de similitudes entre los diferentes pasos que los componen [Marbán et al., 2009]. Entre los modelos disponibles en la literatura, se opta por el uso de un modelo híbrido definido en [Cios et al., 2010], ya que representa un equilibrio entre la rigurosidad técnica de los modelos académicos y la efectividad propia de los modelos industriales. Otra bondad a mencionar de este modelo híbrido es que mantiene un número no muy elevado de pasos lo cual contribuye al potencial del KDP notoriamente.

En la figura 3.1 se muestran las diferentes fases que contempladas en el modelo híbrido seleccionado, así como la correspondencia con las secciones de este capítulo.

3.1.2. Objetivos

Los resultados obtenidos a lo largo del capítulo se utilizan posteriormente para la caracterización de usuarios desde el punto de vista de la red de acceso y poder realizar una estimación de la demanda de tráfico de Internet. A pesar de que los métodos aplicados en este capítulo nos permiten caracterizar a los usuarios sin importar su ubicación de conexión, debido al objetivo anterior, esta tesis doctoral se centra en aquellas zonas de concentración o agregación en las redes de acceso a Internet residenciales.

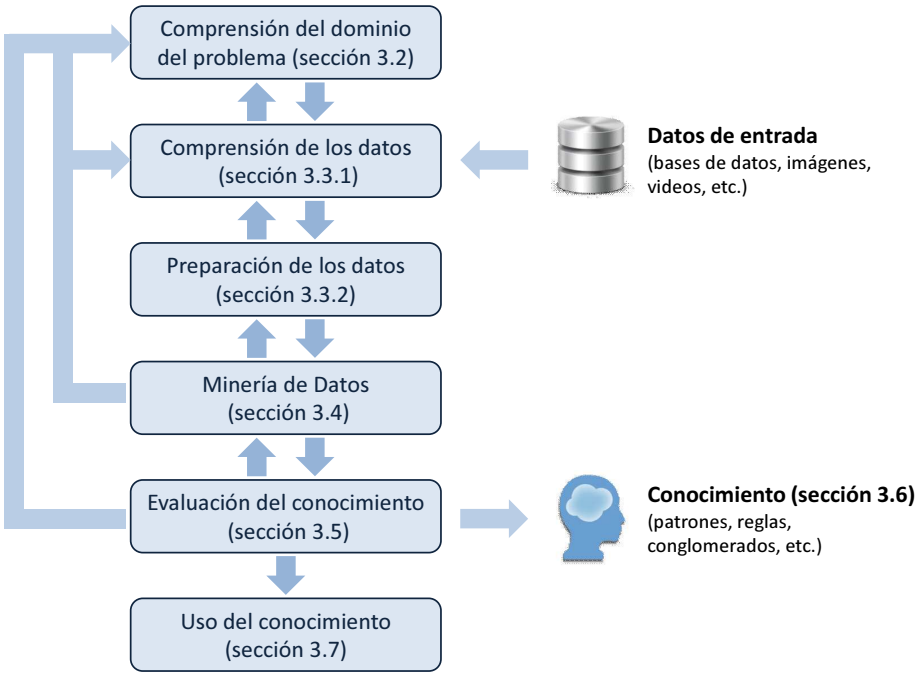

Figura 3.1: Modelo KDP propuesto en [Cios et al., 2010]

Teniendo en cuenta que el concepto de servicio de Internet no tiene por qué ser el mismo desde el punto de vista del usuario y desde el punto de vista de la red, se pretende analizar con especial detenimiento aquellos datos que puedan ser utilizados para extraer conclusiones del uso de los servicios de telecomunicaciones desde el punto de vista de la red. Este planteamiento permitirá realizar una clasificación de la demanda de tráficos de los usuarios de Internet.

El objetivo principal de la caracterización de perfiles de usuarios de Internet es la identificación de patrones de comportamiento de los usuarios en la red, haciendo especial hincapié en descubrir los hábitos de uso en un conjunto determinado de servicios representativos de Internet. Estos hábitos han de ser representativos en relación con la frecuencia de uso que hacen de los servicios de telecomunicaciones, de forma que posteriormente puedan ser utilizados como parámetros de entrada en otros capítulos de la tesis doctoral.

El objetivo puede desglosarse en un conjunto de objetivos secundarios que pueden ser relacionados, desde el punto de vista conceptual, con las fases anteriormente mencionadas del KDP (figura 3.2):

- Desarrollo de modelo conceptual: antes de comenzar las técnicas de minería de datos es recomendable desarrollar un modelo conceptual sobre qué segmentos de usuarios se pueden a encontrar. Este modelo ha de ser coherente con la literatura

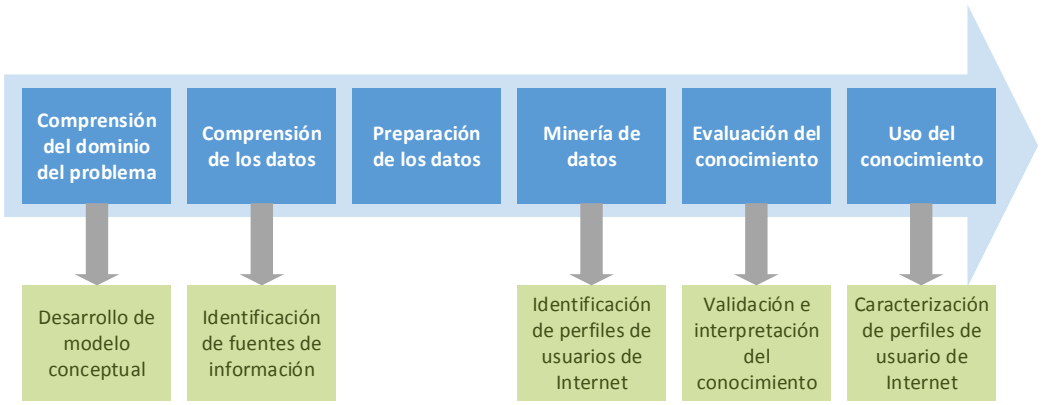

Figura 3.2: Objetivos de las fases del KDP para la caracterización de perfiles de usuario

y se utilizará para su comparación con los segmentos que se encuentren como resultado de la minería de datos.

- Identificación de fuentes de información válidas: se estudian las diferentes fuentes de información disponibles y se evalúa la validez de las mismas para su uso en la caracterización de perfiles de usuario de Internet.
- Identificación de perfiles de usuario de Internet: este objetivo engloba tanto la identificación de aquellos individuos que no utilizan Internet, como la identificación de los diferentes perfiles de usuarios de Internet que existen.
- Validar e interpretar el conocimiento extraído: con el conocimiento extraído de las fases anteriores se podrá validar la segmentación realizada mediante el estudio de métricas de validez. Además, gracias a la caracterización de los perfiles se podrá interpretar el conocimiento extraído.
- Caracterización de los perfiles de usuario: una vez identificado los perfiles de usuario de Internet, se caracterizarán estos perfiles con el fin de conocer sus principales características demográficas.

3.2. Comprensión del dominio del problema

Debido a la complejidad asociada al análisis de conglomerados, es recomendable el desarrollo de un modelo conceptual preliminar a la aplicación de técnicas de minería de datos. De esta forma, se realiza un estudio que ayudará a la comprensión de los resultados extraídos posteriormente a lo largo del KDP.

En esta sección se extraen, a partir del análisis de estado del arte, las características de mayor relevancia en los modelos de usuarios de Internet y que por lo tanto tienen un gran impacto en la clasificación de los distintos perfiles de usuarios que puedan existir. A partir de las experiencias y conocimientos extraídos de la bibliografía consultada, este modelo conceptual define formalmente los factores que influyen en la extracción de perfiles de usuarios y qué perfiles tienen tendencia a aparecer en estudios similares.

El proceso de desarrollo del modelo conceptual puede descomponerse en las siguientes fases:

- Identificación de variables para caracterizar a los segmentos de usuarios de Internet
- Identificación de factores de mayor relevancia en los que se encuentran basados otros estudios
- Identificación de tipologías observables en otros estudios de la literatura

El modelo conceptual consiste, en primera instancia, en una definición formal de los factores de mayor relevancia para realizar una clasificación de usuarios de Internet. Posteriormente, se identifican las tipologías observables en otros estudios, de forma que se extrae un conjunto de perfiles de usuarios que podrían aparecer como resultado del KDP. Este modelo conceptual se utiliza al final del KDP como un método de validación adicional para evaluar la coherencia con la literatura del conocimiento extraído a partir de las técnicas de minería de datos.

3.2.1. Características y factores de relevancia

En primer lugar se extraen un conjunto de características y factores que son los más influyentes en la identificación y caracterización de los perfiles de usuarios. Esta extracción se realiza a partir de un análisis de los modelos teóricos y estudios descritos en las secciones 2.1.2 y 2.1.3 respectivamente.

A pesar de que los modelos teóricos se basan en características difíciles de cuantificar a mediante datos estadísticos, éstos confirman la existencia de factores determinantes en el comportamiento de un individuo en relación al consumo de servicios de Internet, y que por lo tanto, tienen una repercusión directa en la caracterización de perfiles de usuarios de Internet.

En [Venkatesh et al., 2012] se confirma la existencia de variables externas que influyen directamente en la adopción y uso de las TICs. Estas variables son heterogéneas y pueden clasificarse en diferentes categorías:

- Atributos personales: edad y género.
- Status social: educación recibida, ingresos mensuales, lugar de nacimiento.

- Características tecnológicas: experiencia adquirida, utilidad percibida, facilidad de uso, coste de acceso, etc.
- Configuración socio-cultural: otros factores como por ejemplo la influencia social para la adopción tecnológica.

Estas variables influyen directamente en la adopción de la tecnología, siendo críticas a la hora de que un individuo sea un usuario de Internet. En esta adopción tecnológica se podría realizar una caracterización de los usuarios de Internet siguiendo el modelo de difusión de la innovación [Rogers, 2010].

Además, varios estudios sobre usuarios de Internet se basan en los patrones de uso con los que los individuos utilizan los servicios [Selwyn et al., 2005, Howard et al., 2001, ONTSI, 2006]. Los patrones de uso de servicios de Internet se pueden definir en base a los siguientes dos componentes:

- la variedad de servicios de Internet utilizados
- la intensidad en el uso de los servicios de Internet

Por último, a partir de los modelos teóricos y estudios de tipologías de usuario de Internet, se puede concluir que los diferentes usos de servicios tienen un impacto directo en muchos ámbitos de la vida de los usuarios y que podrían verse en forma de gratificaciones o motivaciones. A partir de los modelos teóricos y algunos estudios de usuarios de Internet [McQuail, 2010, Horrigan, 2009, Johnson and Kulpa, 2007] se extraen las siguientes gratificaciones o motivaciones:

1. Información: actividades que aportan un conocimiento (búsqueda de información, foros y blogs).
2. Entretenimiento: actividades accesorias y de ocio (juegos en red, televisión por Internet).
3. Interacción social: actividades de comunicación con otros individuos (redes sociales, chats).
4. Utilidad: actividades que aportan una utilidad y valor añadido al usuario (trabajo desde casa, comercio electrónico).

Destacar la existencia de estudios que tienen en cuenta los dos últimos enfoques descritos. Por un lado, analizan la variedad y frecuencia de uso de los servicios de Internet, y por otro, la utilidad o motivación que se encuentra detrás del uso de los mismos. Este es el caso de los estudios descritos en [Brandtzæg et al., 2011, Brandtzæg, 2010, Horrigan, 2007, Ortega Egea et al., 2006].

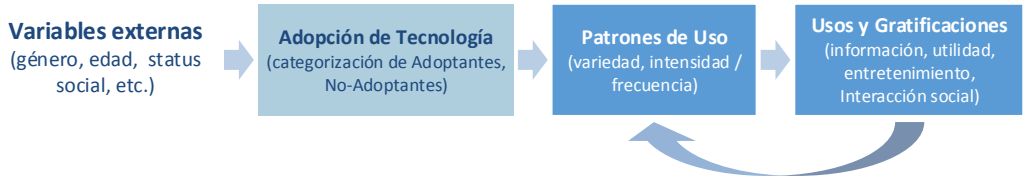

Figura 3.3: Proceso en cadena para la adopción de uso y gratificaciones de Internet

Todos las características y factores de relevancia pueden verse como parte de un proceso que define la adopción, uso, gratificaciones y motivaciones asociadas al consumo de servicios de Internet desde el punto de vista de los usuarios [Fan-Bin Zeng, 2011]. En la figura 3.3 se muestra el proceso compuesto por componentes que se corresponden a las características y factores extraídos de modelos teóricos y estudios de la literatura:

1. Las variables externas influyen en la decisión de un individuo en adoptar o no la tecnología, en este caso el acceso a Internet.

2. Cada usuario de Internet se encuentra caracterizado por un patrón de uso, que se descompone en variedad de servicios utilizados y en frecuencia de uso de los mismos.

3. El consumo de los servicios de Internet tiene un impacto directo en los tipos de usos y gratificaciones percibidas por el usuario. El tipo de uso y de las gratificaciones del usuario influyen, a su vez, en el patrón de uso de los usuarios.

3.2.2. Tipologías observables: modelo conceptual

Las tipologías de usuarios de Internet se centran habitualmente en los últimos niveles del proceso en cadena descrito anteriormente, es decir, en los patrones de uso y gratificaciones percibidas por el usuario de Internet.

La adopción de tecnología, en la que se identifica si un usuario utiliza o no Internet, se obtiene en la gran mayoría de estudios de forma inmediata. Por ejemplo, los estudios que se basan en estadísticas utilizan cuestiones específicas en las que se pregunta al individuo si ha utilizado algún servicio de Internet durante un periodo determinado (habitualmente 12 meses).

Por consiguiente, el modelo conceptual se fundamenta en los parámetros más relevantes para la identificación de perfiles de usuario de Internet, es decir, en los patrones de uso de los servicios y en las gratificaciones percibidas por los usuarios. En la figura 3.4 se muestran las tres dimensiones del modelo conceptual para caracterizar a los perfiles

Figura 3.4: Dimensiones de modelo conceptual de perfil de usuario de Internet

de usuarios de Internet. Estas dimensiones serán utilizadas en la fase de minería de datos.

A partir del análisis de los diferentes estudios de la literatura que siguen enfoques similares, se observa que existen algunos perfiles de usuarios que se describen reiteradamente, si bien a veces se encuentran etiquetados con nombres distintos:

- No-Usuarios: individuos que nunca utilizan Internet desde el hogar.
- Usuarios Esporádicos: usuarios que utilizan pocos servicios de Internet y con una intensidad baja. La frecuencia con la que se conectan a la red es también baja.
- Usuarios Instrumentales: usuarios cuya principal característica es que utilizan los servicios de Internet con una finalidad asociada a una utilidad determinada. En principio, se presupone que la variedad y frecuencia de uso de servicios es moderada.
- Usuarios Sociales: usuarios cuya finalidad principal es la comunicación con otros individuos, tanto desde un punto de vista instrumental como de ocio. El principal servicio que usan son las redes sociales y en segundo plano, foros y blogs. La variedad de servicios y su frecuencia puede ser variada.
- Usuarios de Entretenimiento: usuarios cuya características principal es que utilizan la red para actividades lúdicas y de ocio. Claros ejemplos de este tipo de servicios son la visualización de televisión por Internet, los juegos en red, escuchar música, etc. La variedad y frecuencia de uso de servicios puede ser variada.
- Usuarios Avanzados: usuarios con un uso intensivo y variado de servicios de Internet. Para estos usuarios, Internet es fundamental en sus vidas, por lo que la frecuencia de conexión a la red es alta.

Este conjunto de perfiles de usuarios son una clara contribución a la comprensión del dominio del problema, ya que constituyen una clasificación preliminar de los tipos de usuarios que pueden existir en la población. En la fase de *evaluación del conocimiento* del

KDP estos segmentos de usuarios serán utilizados a modo de comparación de aquellos extraídos en la fase de *minería de datos*.

3.2.3. Objetivos del KDP y de minería de datos

Después de comprender el dominio del problema, se definen los objetivos asociados al KDP y, más específicamente, a la minería de datos.

El objetivo principal consiste en identificar los diferentes tipos de usuarios de Internet existentes basándose en el uso que hagan de Internet, es decir, en los indicadores de comportamientos descritos anteriormente. Uno de los perfiles identificados, ha de ser un conjunto de individuos que no usan la red. La tipología resultante para el resto de usuarios de Internet, proveerá una mejor comprensión de cómo la población de Internet hace uso de la red y consume sus servicios.

El segundo objetivo identificado para este estudio es la caracterización de la tipología de usuarios de Internet en términos de indicadores socio-demográficos, como por ejemplo, género, edad, nivel educacional, etc. Esta caracterización puede utilizarse para describir las relaciones entre segmentos de usuarios y las variables socio-demográficas.

Los objetivos asociados a la fase de minería de datos del KDP pueden ser definidos directamente a partir de los objetivos anteriores. La clasificación de usuarios de Internet será realizada mediante el uso de técnicas de análisis de conglomerados. La caracterización socio-demográfica no requiere una técnica específica de minería de datos.

El alcance de este KDP se encuentra limitado a los usuarios residenciales en España, los cuales consumen servicios de Internet desde sus hogares y mediante tecnologías fijas de red de acceso.

3.3. Comprensión de los datos

Antes de proceder con esta fase del KDP, se ha de analizar e identificar la fuente de información que va a ser utilizada como datos de entrada. Para esta tarea es necesario analizar las diferentes fuentes teniendo en cuenta diferentes parámetros de calidad. Este paso requiere por tanto un análisis exhaustivo de las fuentes de información, describiendo el alcance y la información disponible.

En la sección 2.1.4 del capítulo anterior se presentaron y detallaron las fuentes de información más relevantes para extraer el conocimiento para la segmentación de usuarios de Internet. A partir de ellas, es recomendable realizar un análisis de la calidad de los datos disponibles para seleccionar aquella fuente más adecuada, con objeto de maximizar la calidad de los resultados del análisis de conglomerados que será realizado en la fase de *minería de datos* del KDP [Sivogolovko, 2013].

Para este análisis de calidad se consideran las siguientes dimensiones:

- Validez: los datos han de ser lo suficientemente correctos para representar a todas las entidades de la realidad que se quiere modelar.
- Precisión: conformidad entre los atributos de los datos y las entidades de la realidad que se modela.
- Completitud: grado en el que los datos (o muestras) representan la totalidad de la realidad a modelar.
- Consistencia: las muestras han de ser uniformes en términos de formato y definición a lo largo de todo el sistema o universo bajo estudio. La consistencia asegura que no existan conflictos o contradicciones en un mismo conjunto de datos.
- Validez temporal: Los rangos temporales utilizados en los datos han de ser consistentes y coherentes entre sí para que se represente una realidad específica y en un periodo determinado.

Las fuentes de información han de contener suficiente información sobre el consumo de servicios de Internet y los hábitos de los usuarios para representar la realidad que se quiere modelar mediante el KDP. Este requisito se corresponde directamente con la dimensión de validez. Las 4 dimensiones de calidad restantes han de ser examinadas con detalle para cada una de las fuentes de información disponibles.

Las fuentes de información, descritas en la sección 2.1.4, que cumplen el requisito de validez son las siguientes:

- INE: *Encuesta sobre el equipamiento y uso de tecnologías de la información y comunicación en los hogares españoles* [INE, 2012].
- AIMC: *Estudio General de Medios (EGM)* [AIMC, 2013a].
- CIS: *Barómetro de Junio 2012* [CIS, 2012].
- ONTSI: *Las TIC en los hogares españoles. Datos de actitudes, usos equipamiento y gasto TIC.* [ONTSI, 2013b].

A continuación, se comparan las 4 restantes dimensiones de calidad para cada una de las fuentes de información. Los estudios con mayor precisión son los realizados por AIMC y CIS respectivamente, ya que contienen datos muy específicos sobre el uso de servicios de Internet. Las 4 fuentes presentan una buena calidad en cuanto a las dimensiones de completitud y consistencia. La validez temporal es una desventaja para el estudio realizado por el CIS debido a la periodicidad mensual de sus barómetros. Además, un análisis de la evolución del uso de Internet a partir de este estudio no es posible, ya que las preguntas del estudio varían cada mes. En la tabla 3.1 se plasman visualmente las diferencias entre fuentes de información.

Estudio	Validez	Precisión	Completitud	Consistencia	Validez Temporal
INE	+	–	+	+	+
AIMC	+	+	+	+	+
CIS	+	+	+	+	–
ONTSI	+	–	+	+	+

Tabla 3.1: Comparativa de calidad entre fuentes de información

La fuente de información seleccionada para ser utilizada como dato de entrada en el KDP es la que se corresponde con el estudio realizado por el AIMC y se describe en la siguiente sección.

3.3.1. Estudio General de Medios (EGM)

El EGM nace en 1968 promovido por un grupo de empresas que pretendía consolidar de forma definitiva el estudio de audiencias. Este trabajo consistía inicialmente en 8 oleadas de 4.000 entrevistas, cada una para un periodo de dos años. En el año 1988 nace la asociación AIMC, productora y propietaria del EGM, la cual fomenta el estudio con objeto de investigar las audiencias de los diferentes medios de comunicación y distribuir los resultados entre los miembros de la asociación. A lo largo de más de cuarenta años, el EGM ha sufrido numerosos cambios y avances en su metodología.

3.3.1.1. Características técnicas del EGM

El EGM es un estudio poblacional, donde el principal objetivo es representar de forma adecuada a la población de España, respecto a su comportamiento relativo al consumo de medios. Este estudio tiene una periodicidad anual, aunque se ha diseñado de forma que el ciclo muestral se completa en tres oleadas. Además, el EGM está considerado uno de los estudios con mayor tamaño muestral entre los estudios de audiencia que se realizan en el mundo. Las características técnicas del estudio se encuentran descritas en la tabla 2.3 del capítulo anterior. Los datos seleccionados para la aplicación del KDP son los correspondientes al año más reciente disponible durante la elaboración de esta tesis doctoral, que se corresponde con la anualidad 2012.

El EGM se elabora a partir de un cuestionario con numerosos temas, entre los que destacan por su interés para la extracción de perfiles de usuarios de Internet y su caracterización socio-demográfica, los siguientes:

- **Socio-demográfico y socio-económico.** En esta parte del estudio se define a la población desde un punto de vista socio-demográfico y socio-económico. Algunos de los parámetros presentes en este apartado son los siguientes: edad, sexo, ocupación, nivel de ingresos, hábitat, nivel de estudios, etc. Estos datos son muy importantes

para la caracterización de los diferentes perfiles de usuarios de Internet que se extraerán durante este capítulo y que posteriormente serán utilizados en el modelo predictivo.

- **Medios.** Esta parte del cuestionario comprende varias subcategorías, donde se recogen datos sobre: Prensa Diaria, Suplementos, Revistas, Cine, Radio, Televisión, Internet, y Exterior. Es esta parte del cuestionario la de especial interés para la caracterización de usuarios de Internet, pues se recogen preguntas precisas sobre los hábitos de consumo de servicios de Internet por parte de la audiencia del estudio. La información recogida en estos medios responde al siguiente esquema:

  - Audiencia del último periodo: el valor del periodo corresponde al de la última aparición del soporte ("ayer" para el caso de diarios, radio, televisión, Internet y exterior; y periodos superiores para el cine, suplementos, revistas semanales, quincenales o mensuales).
  - Hábitos de audiencia: se recoge la frecuencia habitual del usuario con cada soporte.
  - Cualificación de la audiencia: se obtiene a través de informaciones complementarias al soporte, como por ejemplo, cantidad leída, lugar o modo de lectura, tiempo dedicado, etc.

- **Equipamiento del hogar.** Esta parte tiene como objetivo la recolección de datos relacionados con el equipamiento general del hogar. Esta información, aunque no se encuentra directamente relacionada con el ámbito de esta tesis doctoral, puede ser utilizada para caracterizar el nivel de equipamiento de los hogares de los usuarios de Internet.

Con objeto de resaltar el volumen de los datos disponibles en el EGM y la complejidad asociada a su tratamiento, destacar que el tamaño muestral, correspondiente al año 2012, cuenta con casi 150.000 muestras de individuos. Además, considerando estos tres temas de interés, se dispone de aproximadamente 80 variables que pueden ser utilizadas para la caracterización de usuarios de Internet.

3.3.2. Preparación de los datos

La preparación de los datos es un paso esencial en el KDP, ya que éstos constituyen los parámetros de entrada para las técnicas de minería de datos. Este paso puede ser dividido a su vez en cuatro diferentes tareas: selección de datos, limpieza de datos, transformación de datos y reducción de dimensionalidad.

3.3.2.1. Selección de datos

En primer lugar, y como parte del proceso de preparación de datos, se han de seleccionar aquellas variables de especial interés para extraer el conocimiento en el KDP. Siguiendo con los temas presentes en el EGM, se seleccionan las siguientes variables divididas en tres categorías correspondientes a los temas anteriormente citados.

La primera categoría cuenta con todas aquellas variables que pueden ser útiles a la hora de caracterizar a los usuarios desde una perspectiva socio-demográfica:

- Datos geográficos: provincia, municipio, hábitat.
- Datos del hogar: individuos en el hogar.
- Datos entrevistado: sexo, edad, estado civil, rol familiar, parentesco, situación laboral, estudios.
- Otros datos entrevistado: nacionalidad, ingresos en el hogar.

En la siguiente categoría se encuentran aquellas variables que describen el uso que hacen los consumidores de los servicios de Internet y sus hábitos más importantes. Además también existen otras variables que pueden ser de gran utilidad como, por ejemplo, características técnicas sobre la conexión a Internet. Destacan los datos sobre la frecuencia y última conexión a los diferentes servicios que se incluyen en el cuestionario.

- Uso de Internet: última vez y frecuencia.
- Equipo de acceso a Internet.
- Tipo de acceso a Internet.
- Proveedor de acceso a Internet.
- Servicios de Internet utilizados en el último año: tipo de servicio, última vez y frecuencia de uso.
- Tiempo de conexión a Internet: lugar de conexión, horas y minutos.
- Último acceso a Internet.

Finalmente, en la última categoría se recogen algunas cuestiones referentes a las características de equipamiento de las casas en relación con el acceso a Internet. Estas variables pueden ser utilizadas para realizar filtros y para caracterizar la antigüedad de los usuarios de Internet.

- Equipos con conexión a Internet.
- Equipamiento informático: ordenadores, portátiles, tablets, acceso de Internet en el hogar, desde cuando tiene acceso a Internet.

3.3.2.2. Limpieza de datos: valores atípicos y perdidos

Esta tarea consiste principalmente en analizar los datos para realizar una limpieza de las variables que constituyen los datos, como por ejemplo mediante el análisis del ruido, o la identificación de valores perdidos y/o atípicos. Además, es importante que los datos de entrada al KDP satisfagan ciertas condiciones o requisitos impuestos por las técnicas de minería de datos. En caso de que existen estas condiciones, es posible que se requiera una realimentación en el KDP.

En el caso de los datos seleccionados para la caracterización de usuarios en Internet, no hace falta realizar estas acciones debido a que las muestras con valores atípicos corresponden a individuos con comportamiento atípico y éstos han de estar representados en los perfiles de usuarios de Internet extraídos.

En relación a los valores perdidos, aquellas variables que describen frecuencia de uso de servicios, se considera que el individuo no realiza ningún uso del servicio. En cuanto a otras variables, sobretodo de tipo socio-demográfico, los valores perdidos no se reemplazan por otros valores debido a que no pueden ser predichos.

3.3.2.3. Transformación de datos

En esta tarea se engloba cualquier proceso en los que se definen nuevos conjuntos de variables a partir de transformaciones, derivaciones o cambios de tipo o rango. En muchas ocasiones, esta transformación de variables de los datos se realiza para satisfacer requisitos propios de las técnicas de minería de datos. Si se cambiase de técnica de minería de datos es muy probable que se requiera una realimentación en el KDP.

En la colección de datos seleccionada se requiere la realización de dos transformaciones fundamentales en dos conjuntos de variables distintos:

- Numerización y normalización de variables de frecuencia de uso de servicios de Internet
- Discretización de variables socio-demográficas

Numerización y normalización de variables de frecuencia de uso. Las variables que indican frecuencia de uso de servicios de Internet se encuentran representadas con una escala de tipo *likert* u ordinal [Likert, 1932]. Este tipo de escala son nominales pero con un significado en orden creciente o decreciente. Un ejemplo habitual es una variable que mide la satisfacción de un usuario: muy (1), bastante (2), medio (3), poco (4) y muy poco (5).

El problema asociado con este tipo de variables es que la distancia entre valores de la escala no tiene por qué ser lineal con los valores conceptuales que representan.

Este fenómeno afecta negativamente en la aplicación de algunas técnicas de análisis de conglomerados, ya que se utilizan métricas basadas en distancias.

Este es el caso de las respuestas a las cuestiones referentes a la frecuencia de uso de los servicios de Internet. Las respuestas tienen un valor numérico que describe en orden decreciente la frecuencia de uso de un servicio, en este caso definiendo 5 niveles. El objetivo de esta transformación es la conversión en variables de tipo escalar y que representen directamente el número de veces de uso de un servicio (días) en un periodo de tiempo determinado y en una misma escala temporal. Para realizar esta transformación se traducen los valores descriptivos en valores escalares siendo fiel a la descripción de los niveles anteriormente descritos:

- Casi todos los días (1) $\rightarrow$ 4-7 por semana
- Dos o tres veces a la semana (2) $\rightarrow$ 2-3 por semana
- Una vez por semana (3) $\rightarrow$ 1 por semana
- Una ó dos veces al mes (4) $\rightarrow$ 1 a 2 por mes
- Menos de una vez al mes (5) $\rightarrow$ $<$ 1 por mes

A continuación se ajustan los niveles para que éstos se encuentren definidos en una misma escala temporal. La escala utilizada es la de frecuencia de uso a la semana, partiendo de que 1 mes tiene de media aproximadamente 4,35 semanas. Además, se ajustan los valores de los niveles de forma que su significado coincida con los valores medios de los rangos que comprendían:

- Casi todos los días: 4-7 por semana $\rightarrow$ 5.5 por semana
- Dos o tres veces a la semana: 2-3 por semana $\rightarrow$ 2.5 por semana
- Una vez por semana: 1 por semana $\rightarrow$ 1 por semana
- Una o dos veces al mes: 1-2 / 4.35 por semana $\rightarrow$ 0.35 por semana
- Menos de una vez al mes: Aproximadamente 0 por semana $\rightarrow \approx 0$ por semana

Discretización de variables. Existe un conjunto de variables socio-demográficas, que aunque no son parámetros de entrada para la fase de minería de datos del KDP, se utilizan para una posterior caracterización de los perfiles de usuario de Internet. El objetivo de la discretización es la de facilitar la interpretación de las variables numéricas convirtiéndolas en nominales o escalares. Para las variables que indican la edad y el número de miembros en el hogar, se ha realizado una discretización de *1 a 1*, donde se han utilizado los intervalos ya tipificados por el INE y EUROSTAT. A continuación, se presentan los intervalos utilizados para las dos variables anteriores:

- Edad: 14-19 / 20-24 / 25-34 / 35-44 / 45-54 / 55-64 / 65 y más años
- Tamaño de hogar: 1 / 2 / 3 / 4 o más

Otras transformaciones de variables. Se ha realizado otra transformación para crear una nueva variable a partir de la información de otras variables y que indique qué tipo de acceso a Internet tienen los diferentes usuarios. Las variables que contienen la información de la nueva variable se corresponden con las siguientes cuestiones:

- Acceso a Internet desde el hogar
- Acceso a Internet a través de ADSL
- Acceso a Internet a través de Cable
- Acceso a Internet a través de un dispositivo móvil
- Acceso a Internet a través de un dispositivo USB
- Acceso a Internet a través de WIFI de acceso gratuito
- Acceso a Internet a través de otros sistemas

Así pues, la nueva variable se define en cinco niveles de acceso a Internet, los cuales se calculan mediante las anteriores variables:

- Sin acceso (1): el individuo no dispone de acceso a Internet en el hogar
- Acceso móvil a Internet (2): el usuario sólo accede a la red a través de WiFis gratuitas o a través de redes móviles. No dispone de acceso a Internet a través de ADSL ni cable.
- Acceso a Internet a través de ADSL (3): el usuario se conecta a Internet a través de una red de acceso fija, tradicionalmente de tipo xDSL.
- Acceso a Internet a través de cable (4): el usuario se conecta a Internet a través de una red de acceso de cable coaxial.
- Acceso a Internet a través de otros sistemas (5): el usuario accede a Internet a través de otros sistemas no especificados.

Por último también destacar que no aparecen los accesos a Internet a través de otros sistemas que coexisten en el territorio español, como por ejemplo la fibra óptica. En este caso, los usuarios de estos sistemas, se encontrarán bajo la categoría de *Acceso a Internet a través de otros sistemas (5)* o bien bajo la categoría de *Acceso a Internet a través de ADSL (3)*. La causa más probable es que en la actualidad los términos de

Internet y ADSL se utilizan de forma indistinta. Por esta razón, esta variable ha de ser considerada con cierta cautela debido a que los casos que describe pueden ser imprecisos en cuanto a la tecnología de red de acceso del usuario.

3.3.2.4. Reducción de dimensionalidad

Con el objetivo de reducir las dimensiones de los datos necesarios para realizar la identificación de perfiles de usuarios de Internet, se procede a la selección de un subconjunto adecuado de variables que permitan esta extracción del conocimiento mediante técnicas de minería de datos.

La colección de datos disponible contiene 14 variables distintas que indican la frecuencia de uso de servicios de Internet, lo cual supone numerosas posibilidades de combinaciones de subconjuntos (ecuación 3.1).

$$N_{comb} = \sum_{k=1}^{13} \binom{13}{k} = \sum_{k=1}^{13} \frac{13!}{k!(13-k)!} = 8191 \tag{3.1}$$

Por esta razón una estrategia de búsqueda de subconjuntos completa es inabordable, por lo que se ha optado por una estrategia heurística. Además, la dirección seleccionada ha sido *hacia atrás*, ya que se parte de la totalidad de las variables y se han ido filtrando a la vez que se analizan los resultados obtenidos de las técnicas de minería de datos. Esta dirección favorece la extracción de un modelo que cuente con la mayor información posible para identificar los perfiles de usuario de Internet.

Para seleccionar aquellas variables con mayor relevancia en la extracción del conocimiento, se parte del modelo conceptual desarrollado en la sección 3.2. Además, se emplean las técnicas descriptivas descritas en la sección 2.1.6.1 (análisis factorial y análisis de componentes principales) para contribuir a una mejor comprensión de la información aportada por cada variable. También se da mayor importancia a aquellas variables que puedan tener un mayor impacto en el consumo de recursos de red, tal y como se describe en la sección 2.2.4.

Basándose en estas dos premisas se considera, en primera instancia, que las siguientes variables son las más relevantes:

- Compartición de Archivos
- Redes sociales
- Lectura de Información
- Programas y Series de TV
- Operaciones con Entidad Bancaria

- Compra de Productos y Servicios

Después de haber realizado varias iteraciones en el KDP, es decir, aplicar las técnicas de minería de datos con varios subconjuntos de variables, se concluye que otras variables tienen también relativa relevancia para dotar de calidad al modelo extraído, éstas son:

- Correo Electrónico: parece un buen indicador de uso general de la red y de adopción de la tecnología.

- Juegos en Red: servicio consumidos por usuarios con perfiles más avanzados y que disponen de mayor adopción tecnológica.

Sin embargo, a lo largo de varias iteraciones donde se han analizado la exclusión de diferentes variables, se ha comprobado que algunas de éstas no contribuyen al diseño de modelo de calidad. En algunos casos, estas variables han sido filtradas debido a que el propio significado de las mismas era poca preciso y confuso. Este es el caso de las siguientes variables:

- Mensajería Instantánea: esta característica incluye también servicios de mensajería instantánea en dispositivos móviles (por ejemplo, *WhatsApp*). Por esta razón, esta variable es poco precisa en cuanto al uso doméstico que hacen los usuarios del servicio.

- Telefonía sobre Internet: engloba cualquier servicio de telefonía sobre Internet (como por ejemplo, *Skype*), así como los servicios desplegados por operadores de telefonía fija sobre sus redes de acceso. No obstante, muchos usuarios desconocen que la telefonía fija disponible en sus hogares es de este tipo, por lo que las respuestas a esta cuestión pueden ser imprecisas.

En otros casos, algunas variables indicaban usos de servicios muy específicos que no concuerdan con el modelo conceptual descrito en la sección 3.2. En algunos casos, la inclusión de estas variables ha conllevado a que el modelo extraído se encuentre polarizado por los servicios asociados a las mismas, perdiendo por tanto calidad. Este ha sido el caso de las siguientes variables:

- Blogs y Foros: su uso se encuentra muy extendido por lo que no es suficientemente representativo. Además, el análisis de frecuencias de esta variable podría indicar que muchos usuarios visitan blogs y foros y no son conscientes de ello.

- Podcasts: los casos que hacen uso de este servicio son muy escasos, por lo que esta variable es irrelevante para la extracción de conocimiento coherente con el modelo conceptual.

- Escuchar Música: esta variable se encuentra íntimamente ligada al servicio de visionado de series y televisión, y al uso de redes sociales. Su inclusión provoca que exista un gran peso hacia esos servicios, haciendo que los perfiles extraídos del KDP pierdan calidad.

El método de evaluación del modelo extraído a partir de un subconjunto de variables se basa en métricas de calidad objetivas (sección 2.1.7), la coherencia y la adecuación con el modelo conceptual (sección 3.2). La evaluación del conocimiento extraído se describe con más detenimiento en la sección 3.5.

3.4. Minería de datos

Esta fase del KDP tiene como principal objetivo la extracción del conocimiento, en este caso, la tipología de usuarios de Internet. Para el descubrimiento de este conocimiento se aplican ciertos tipos de técnicas de minería de datos, seleccionadas en la sección 3.2.3 de este capítulo.

Para la aplicación del KDP a las tipologías de usuarios de Internet, se propone dividir la fase de minería de datos en tareas secuenciales:

1. Identificar a los individuos que no utilizan Internet y que conforman el segmento de *No-Usuarios*

2. Realizar un análisis descriptivo del segmento de *Usuarios de Internet* para facilitar la comprensión de las variables a utilizar en el análisis de conglomerados

3. Realizar un análisis de conglomerados para identificar y extraer los perfiles de usuarios de Internet

Destacar que la segunda tarea supone una iteración en el KDP, pues la información estadística extraída de este análisis es utilizada para reducir la dimensionalidad y complejidad asociada en los datos (sección 3.3.2.4), así como, facilitar la posterior tarea del análisis de conglomerados.

3.4.1. Identificación de No-Usuarios

Esta sección identifica al segmento de los *No-Usuarios*, compuesto por aquellos individuos que no usan Internet, independientemente de que dispongan de acceso a Internet en sus hogares o no. El segmento complementario, de *Usuarios de Internet*, se puede obtener de forma directa a partir de esta identificación.

La identificación de este segmento de individuos se puede realizar a partir de técnicas de minería de datos de forma conjunta a la identificación de los perfiles de *Usuarios de Internet*. No obstante, se propone la identificación de forma directa, a partir del análisis

estadístico de las variables disponible en la colección de datos, para identificar de forma precisa si un individuo utiliza o no Internet.

La mayor ventaja de este enfoque es que la complejidad, asociada a este paso del KDP, se ve reducida de forma drástica, ya que se reduce en una unidad los segmentos a extraer mediante algoritmos de minería de datos. Además, el tamaño de la muestra que se utiliza para los algoritmos también se reduce, pues se filtran aquellos individuos que no usan Internet.

En la sección anterior se presentan una serie de variables, a partir de las cuales se puede realizar una clasificación para extraer el segmento de *No-Usuarios*:

- En los últimos doce meses, personalmente, ¿Ha accedido / usado alguna vez Internet?
- ¿Con qué frecuencia usa Internet?
- ¿Dispone de acceso / conexión a Internet en el hogar (independientemente de que Vd. lo use o no)?

Las tablas 3.2 y 3.3 muestra las tablas de contingencia de las dos primera variables respecto a la tercera. A partir de esta información se puede extraer de forma directa el segmento de los *No-Usuarios* con diferentes umbrales de tiempo para determinar si un individuo es usuario o no de Internet.

En la primera tabla 3.2 se identifica fácilmente a los usuarios de Internet como aquellos individuos que acceden a la red desde sus hogares durante los últimos 12 meses (53,2 %). El segmento complementario, es decir, el resto de individuos conforman a los *No-Usuarios* ($100\,\% - 53,2\,\% = 46,8\,\%$). Analizando con más detalle este segmento, se aprecia que los *No-Usuarios* se encuentra formado por:

- Individuos sin acceso a Internet en el hogar y que no lo utilizan (31,7 %)
- Individuos con acceso a Internet en el hogar y que no lo utilizan (7,9 %)
- Individuos sin acceso a Internet en el hogar y que sí utilizan servicios de Internet en el último año (7,2 %)

Esta última categoría, no despreciable, corresponde a un colectivo de la población que hace uso de la red pero no desde el domicilio de residencia, como por ejemplo, desde el centro de estudios, desde el trabajo, o desde casa de amigos o familiares. La identificación de este segmento es interesante y no suele ser identificada en la literatura.

La segunda tabla 3.3 presenta la información más detallada sobre la frecuencia de acceso a Internet. A partir de esta tabla de contingencia se puede extraer que, si se varía el umbral anteriormente utilizado (acceso a Internet en los últimos 12 meses) a *varias veces al mes*, la diferencia del tamaño de segmento variaría menos de un 1 %.

		Internet desde el hogar		
		SI	NO	Total
Acceso a Internet en los últimos 12 meses	SI	53,2 %	7,2 %	60,5 %
	NO	7,9 %	31,7 %	39,5 %
	Total	61,1 %	38,9 %	100 %

Tabla 3.2: Tabla de contingencia de acceso a Internet en el último año desde el hogar

		Internet desde el hogar		
		SI	NO	Total
Frecuencia de uso de Internet	Diaria	43,4 %	2,2 %	45,6 %
	Varias veces por semana	6,7 %	2,3 %	9,0 %
	Varias veces al mes	2,2 %	1,7 %	3,9 %
	Con menor frecuencia	0,9 %	1,0 %	1,9 %
	Nunca	7,9 %	31,7 %	39,6 %
	Total	61,1 %	38,9 %	100 %

Tabla 3.3: Tabla de contingencia de frecuencia de uso de Internet desde el hogar

A pesar de que ambas tablas de contingencia se podrían haber utilizado para la identificación del segmento de *No-Usuarios*, se ha utilizado la primera tabla para extraer este segmento. De esta forma, el umbral temporal utilizado para definir si un individuo es usuario de Internet se define en 12 meses.

Una vez extraído el conjunto de muestras que se corresponden con los *No-Usuarios*, se puede obtiene directamente el conjunto complementario correspondiente a los *Usuarios de Internet*. En las siguientes dos secciones, cuyo objetivo es identificar y extraer los perfiles de los usuarios de Internet, se utilizarán como parámetro de entrada tan solo aquellas muestras correspondientes a este segundo segmento de *Usuarios de Internet*.

3.4.2. Análisis descriptivo de usuarios de Internet

El objetivo de esta sección es entender y comprender las variables que van a ser utilizadas como parámetros de entrada en las técnicas de minería de datos. Las muestras que van a ser utilizadas como datos de entrada son las correspondientes al conjunto de individuos que sí utilizan Internet, identificado en la sección anterior.

A partir del modelo conceptual de la sección 3.2.2, se requiere información que indique qué servicios se utilizan y con qué frecuencia se usan, la cual contribuye a una mejor compresión global sobre la variedad de uso y las preferencias de contenidos por parte de los usuarios de Internet. Esta información se obtiene de las siguientes variables de la colección de datos, especificando las cuestiones y posibles respuestas:

- ¿A qué servicios de Internet ha accedido en los últimos 12 meses?

- Correo electrónico.
- Mensajería instantánea (WhatsApp, Messenger o similar).
- Herramientas de compartición de archivos (tipo eMule, Ares, etc).
- Realiza llamadas telefónicas por Internet (telefonía IP, Skype, etc).
- Jugar en Red.
- Redes sociales.
- Participar en Blogs o foros.
- Lectura de información de actualidad.
- Visionado de programas y series de Televisión / películas.
- Escuchar música directamente en Internet.
- Escucha / descarga de podcast.
- Operaciones con su entidad bancaria.
- Compra de productos y servicios.

- ¿Con qué frecuencia ha usado los servicios?
  - Todos casi todos los días.
  - 2 o 3 veces en la semana.
  - Una vez por semana.
  - Una o dos veces al mes.
  - Menos de una vez al mes.

Se considera que la variable correspondiente a la frecuencia de uso de los servicios es más relevante y precisa, dado que el objetivo es el análisis de los hábitos en el consumo de servicios de Internet.

3.4.2.1. Análisis de frecuencias

El análisis de frecuencias tiene como principal objetivo conocer el número de veces que toman cada uno de los valores de las variables. Con este análisis se puede conocer qué servicios se utilizan con más frecuencia en la población y cuáles menos. De esta forma, también se identifica aquellos servicios más populares y aquellos más específicos. Esta información es vital para conocer el comportamiento de las variables a la hora de ser utilizadas en un algoritmo de conglomerados. Esta información aporta datos para realizar una reducción de la dimensionalidad más eficiente (sección 3.3.2.4).

En la figuras 3.5 y 3.6 se muestran la proporción de los valores que toman todas las variables de frecuencia de uso de los servicios de Internet.

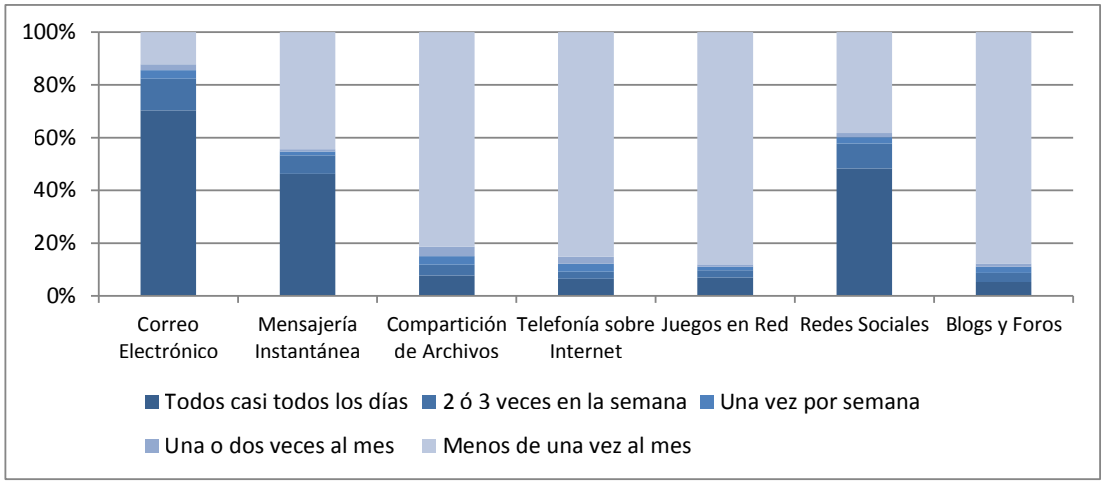

Figura 3.5: Análisis de frecuencias de variables de frecuencia de uso de servicios (I)

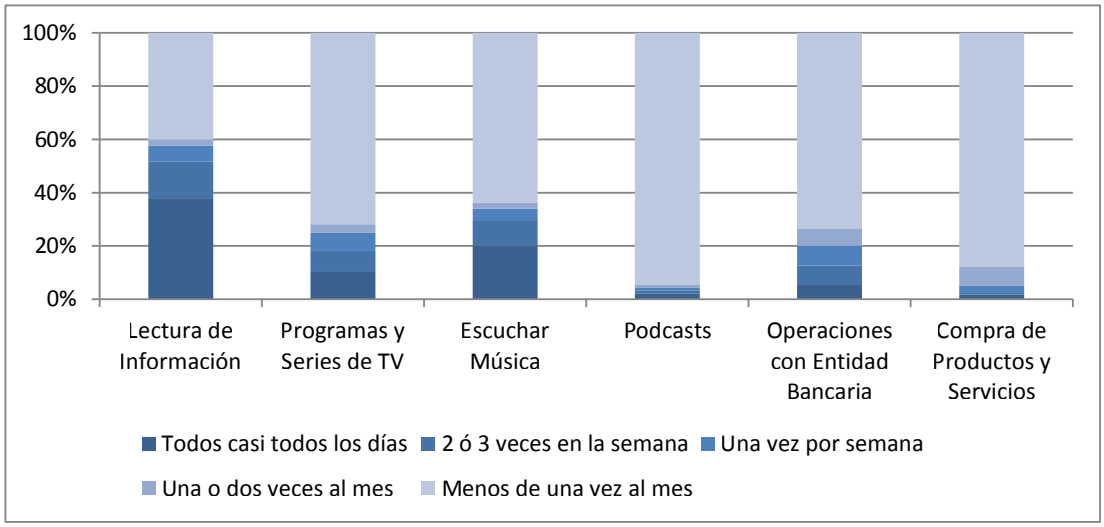

Figura 3.6: Análisis de frecuencias de variables de frecuencia de uso de servicios (II)

Como puede observarse, el servicio más popular, utilizado ampliamente por casi el 90 % de los usuarios es el correo electrónico. Además, otros servicios, como la mensajería instantánea, las redes sociales, y la lectura de información gozan de gran popularidad entre los usuarios de Internet.

El visionado de programas y series de televisión, escuchar música y las operaciones con entidades bancarias son servicios que cuentan con cierta popularidad ya que se encuentran cercanos al 30 % de frecuencia de uso.

El resto de servicios, cuentan con una baja frecuencia entre los individuos, no superior al 20 %. Algunos de estos servicios podrían verse, en función de su naturaleza y significado, como buenos candidatos para identificar a algunos tipos de usuarios.

3.4.2.2. Análisis de correlaciones

A continuación se presenta el análisis de correlaciones entre las variables anteriores, el cual es importante como paso previo a la aplicación de técnicas de minería de datos de extracción de perfiles de usuarios de Internet. Cuando las variables contienen una gran correlación entre sí, puede que se esté contribuyendo a que los algoritmos doten de mayor peso a algún tipo de servicio o factor, obteniendo por tanto resultados confusos y, en muchos casos, polarizados hacia un tipo de actividad.

En términos generales se considera que una correlación r, implica una determinada asociación entre dos variables de la siguiente forma:

- Si $|r| < 0,3 \rightarrow$ la asociación es débil
- Si $0,30 \leq |r| \leq 0,70 \rightarrow$ la asociación es moderada
- Si $|r| > 0,70 \rightarrow$ la asociación es fuerte

En líneas generales no existen correlaciones elevadas entre las variables, pero sí que se da el caso en el que se dan asociaciones moderadas entre los servicios de Internet (correlaciones entre 0,3 y 0,4). Desde el punto de vista conceptual, puede ser bastante lógico debido a que aquellos usuarios con una mayor adopción tecnológica, tenderán a utilizar varios servicios de Internet.

En la tabla 3.4 se describe el grado de asociación entre variables correspondientes al consumo de servicios de Internet, es decir, el resultado al análisis de correlaciones entre variables:

Las correlaciones más significativas son moderadas y muestran qué variables podrían ser agrupadas entre ellas mismas, desde un punto de vista puramente estadístico. Además, estas correlaciones también pueden tener un significado desde el punto de vista conceptual. Las principales conclusiones que se extraen son las siguientes:

Variable	**Asociación moderada**	**Asociación débil**
Correo electrónico	- Mensajería instantánea	- Redes sociales - Lectura de información
Mensajería instantánea	- Correo electrónico - Redes sociales	- Compartición de archivos
Compartición de archivos		- Mensajería instantánea - Telefonía sobre Internet - Programas y series de TV - Escuchar música
Telefonía sobre Internet		- Mensajería instantánea - Compartición de archivos - Blogs y Foros - Programas y series de TV - Escuchar Música
Juegos en red		- Escuchar Música
Redes sociales	- Mensajería instantánea - Escuchar música	- Correo electrónico - Programas y series de TV
Blog y foros		- Telefonía sobre Internet - Programas y series de TV - Podcasts
Lectura de información		- Correo Electrónico - Programas y series de TV - Op. con entidad bancaria
Programas y series de TV	- Escuchar música	- Mensajería instantánea - Compartición de archivos - Telefonía sobre Internet - Redes sociales - Blogs y Foros - Lectura de información
Escuchar música	- Redes sociales - Programas y series de TV	- Mensajería instantánea - Compartición de archivos - Telefonía sobre Internet - Juegos en red - Podcasts
Podcasts		- Blogs y foros - Escuchar música
Operaciones con entidad bancaria	- Compra de productos y servicios	- Correo electrónico - Lectura de información
Compra de productos y servicios	- Op. con entidad bancaria	

Tabla 3.4: Análisis de correlaciones entre variables de frecuencia de uso de servicios de Internet

- Los servicios de naturaleza multimedia, como por ejemplo *programas y series de TV* y *escuchar música*, se encuentran bastante relacionados.
- Los servicios de *operaciones con entidad bancaria* y *compra de productos y servicios*, se encuentran relacionados.
- Los servicios con una finalidad comunicativa se encuentran correlados entre sí. Este es el caso del servicio de *correo electrónico* que se encuentra correlacionado con la *mensajería instantánea*. La *mensajería instantánea* también se encuentra asociada con las *redes sociales*, medio donde también existen chats y servicios de mensajería. Y por último, las *redes sociales* que tienen una asociación moderada con los servicios de *escuchar música*.

Destacar la existencia de un gran número de asociaciones débiles entre servicios de Internet. Una probable causa de este fenómeno, se encuentra ligada al hecho de que aquellos usuarios con una buena adopción tecnológica de Internet, tengan una mayor tendencia a utilizar un conjunto de servicios, sin importar la complejidad asociada a los mismos.

3.4.3. Análisis de conglomerados: Usuarios de Internet

En esta sección describe el proceso de descubrimiento de los conglomerados existentes en los usuarios residenciales de Internet durante el año 2012. Para este propósito, se aplican técnicas de minería de datos para agrupar entidades (usuarios) con características comunes entre sí para formar conglomerados (perfiles o tipos de usuario). Las entidades que conforman un conglomerado han de ser similares entre sí mismas y disimilares con las de otros conglomerados.

A lo largo de la sección, se describe la metodología empleada para la extracción del conocimiento y posteriormente se detalla su aplicación al caso de estudio de los usuarios de Internet residenciales en España.

3.4.3.1. Selección de técnica de análisis de conglomerados

En la primera fase del KDP (sección 3.2.3) se seleccionó el análisis de conglomerados como la técnica de minería de datos para extraer la tipología de usuarios de Internet. En esta sección se selecciona el tipo de algoritmo a utilizar para extraer los perfiles de usuarios de Internet.

En [Estivill-Castro, 2002] se pone de manifiesto que la selección de una técnica específica de minería de datos es una tarea claramente subjetiva que depende de la persona que realiza el análisis y que no se puede afirmar con rotundidad o de forma objetiva que exista una solución o algoritmo correcto. No obstante, en [Milligan, 1996] se presentan un conjunto de aspectos a tener en cuanto a la hora de realizar esta selección:

1. Capacidad de extracción del tipo de conglomerados que se sospecha que existen.
2. Efectividad para extraer los conglomerados.
3. Robustez ante la presencia de errores en los datos.
4. Disponibilidad de aplicaciones software que puedan ejecutar el algoritmo.

El primer aspecto hace referencia a la existencia de un análisis preliminar sobre los conglomerados que se sospecha que existan en los datos. Este hecho se encuentra alineado con la metodología seguida en esta tesis doctoral, pues en la sección 3.2.2 se define un conjunto de perfiles de usuarios que podrían esperarse como resultado de las técnicas de minería de datos. En [Jain, 2010], debido a la inherente dificultad para extraer conocimiento a partir de análisis de conglomerados, se recomienda la realización de análisis preliminares para el desarrollo de técnicas semi-supervisadas que ayuden en la correcta representación de datos y selección de algoritmos.

A continuación se va a seleccionar la técnica de minería de datos para esta fase del KDP siguiendo las diferentes abstracciones y clasificaciones descritas en la sección 2.1.6.2 de esta tesis doctoral.

A pesar de que se ha realizado un análisis preliminar sobre los posibles perfiles de usuario de Internet que pueden existir en los datos bajo estudio, se decide que la técnica de minería de datos sea de tipo no-supervisada. La razón reside en que este conocimiento extraído del análisis preliminar servirá para validar conceptualmente los resultados, así como su coherencia con otros estudios similares.

El modelo de agrupamiento seleccionado para el caso de estudio de esta tesis doctoral es de tipo particional, debido a que el agrupamiento jerárquico tiene una eficiencia muy baja cuando los datos de entrada del algoritmo tienen un tamaño muestral considerable, como es el caso de la colección de datos utilizada en el KDP.

Por último se selecciona el tipo de algoritmo de minería de datos a utilizar. Como se ha mencionado anteriormente, no existen algoritmos que sean objetivamente mejor que otros, por lo que durante el desarrollo de etapa del KDP se realizaron varias iteraciones analizando la viabilidad y calidad de diferentes algoritmos de análisis de conglomerados basados en diferentes enfoques (centroides, distribuciones estadísticas, densidades).

Después de varias ejecuciones con distintos tipos de algoritmos se opta por el uso de un algoritmo basado en centroides, debido a las características de los datos de entrada utilizados para la caracterización de usuarios de Internet. Los algoritmos basados en distribuciones estadísticas fueron descartados debido a su baja escalabilidad con cantidades de datos considerables y la complejidad asociada a los parámetros de entrada de ajuste del algoritmo. En el caso de los algoritmos basados en densidades, también fueron descartados debido a que consideran ruido o valores atípicos a las muestras localizadas en zonas de baja densidad. La principal consecuencia de este

fenómeno es que, debido a que los datos de entrada tienen una alta dispersión entre muestras, este tipo de algoritmos tendrán cierta tendencia a ignorar a un conjunto no despreciable de muestras que representan usuarios de Internet que deberían ser debidamente considerados.

Otros algoritmos considerados. Además de los algoritmos anteriores, también se analizaron otras alternativas para la extracción de los conglomerados que representan los perfiles de usuario de Internet. Entre las alternativas analizadas, destaca el análisis de componentes principales y el análisis de conglomerados bifásico.

El análisis de componentes principales, descrito en la sección 2.1.6.1, tiene como principal objetivo convertir un conjunto de observaciones de variables posiblemente correladas, en un conjunto de valores de variables linealmente incorreladas, conocidas como componentes principales. Una vez realizado este análisis se puede aplicar unos procedimientos adicionales para extraer una clasificación de muestras a partir de las componentes principales calculadas para todos los casos de la muestra. No obstante, los resultados obtenidos mediante esta técnica no son coherentes con el modelo conceptual (sección 3.2.2). La calidad de los resultados es baja, pues los componentes no describen más del 60 % de la muestra.

El análisis de conglomerados bifásico es un algoritmo presente en la herramienta estadística *IBM SPSS* [Zhang et al., 1996, Chiu et al., 2001]. Sus principales características son su escalabilidad frente a grandes cantidades de datos y su capacidad de manejar variables o atributos escalares y categóricas. Se asumen que las variables continuas y las categóricas se distribuyen de forma normal y multinomial respectivamente, y que éstas son independientes. Debido a que estas suposiciones no se cumplen y que existen correlaciones entre atributos de la muestra de datos, los resultados extraídos mediante esta técnica de minería de datos distan del modelo conceptual. Los resultados extraídos a partir de multitud de ejecuciones del algoritmo utilizando diferentes subconjuntos de atributos tienden a tener tamaños muy dispares de conglomerados, dando lugar a conglomerados excesivamente pequeños o grandes. La causa detrás de este comportamiento es que el algoritmo tiende a dar un peso excesivo a algunas variables haciendo que los resultados tiendan a encontrarse *polarizados* hacia un servicio de Internet en particular.

Conclusiones. Los aspectos propuestos por [Milligan, 1996] han sido fundamentales para la selección del algoritmo de análisis de conglomerados. Además de los argumentos anteriormente citados para la selección del tipo de algoritmo basado en agrupación particional, también se ha evaluado la posibilidad de utilizar otros algoritmos en función de su calidad y eficacia para la extracción del conocimiento a partir de los datos, así como su disponibilidad en software o la complejidad asociada para su implementación.

Se selecciona el algoritmo *K-Means*, detallado en la sección 2.1.6.2, porque es el más

adecuado a las características de los datos de entrada debido a su alta escalabilidad y eficiencia para analizar grandes cantidades de datos [Tan et al., 2006]. Con una adecuada selección de parámetros de entrada, este algoritmo representa un equilibrio entre la efectividad para la extracción de los perfiles de usuario de Internet y la disponibilidad de aplicaciones software, sin descuidar en ningún momento, la robustez y calidad de la técnica de minería de datos.

3.4.3.2. Aplicación de técnica de análisis de conglomerados

A continuación, se detalla el proceso que se ha de seguir para descubrir la tipología de usuarios de Internet. El algoritmo *K-Means* requiere que se definan 3 parámetros de entrada (sección 2.1.6.2): número de conglomerados K, centroides iniciales C_i y métrica de similitud utilizada d.

Debido a que el número de conglomerados K es un parámetro de entrada en el algoritmo y no existe un criterio matemático para seleccionar un valor correcto, se siguen las recomendaciones descritas en la literatura [Tibshirani et al., 2001], donde se plantea un enfoque heurístico para abordar esta limitación. Por lo tanto, se ejecuta el algoritmo múltiples veces con distintos valores de K. Después de cada iteración, se calculan métricas de calidad con objeto de poder seleccionar el valor de K que tiene una mayor calidad.

Otro parámetro de entrada del algoritmo que se ha considerar son los centroides iniciales C_i. En este caso, se opta por la selección de centroides debidamente separados entre sí siguiendo las pautas descritas en [SPSS, 2011, Hartigan, 1975]:

1. Selección aleatoria de K casos de la muestra como centroides iniciales.

2. Se prueba el resto de casos de la muestra para comprobar si éstos pueden substituir a los centroides iniciales anteriormente elegidos en función de si se cumplen unas condiciones determinadas (ver pseudocódigo 2):

   a) Si la distancia entre el caso x_j y su centroide más cercano es mayor que la distancia entre los centroides más próximos entre sí (M_m y M_n), entonces x_j reemplaza al centroide más próximo, ya sea M_m o M_n.

   b) Si la distancia entre el caso x_j y el segundo centroide más cercano a x_j es mayor que la menor distancia entre el centroide más cercano y cualquier otro centroide, entonces x_j reemplaza al centroide más cercano.

En referencia a la métrica de calidad, se hace uso de la distancia más popular utilizada con *K-Means* en otros estudios, la distancia euclidea. Por último, siguiendo las recomendaciones de [Mao and Jain, 1996], se preparan los datos de forma adecuada

Pseudocódigo 2 K-Means: Condiciones de substitución de centroide inicial

Notación: M_i, media de conglomerado i
x_j, vector correspondiente a muestra j (representa un usuario)
$d(x_i, xj)$, distancia euclidea entre vectores x_i y x_j
$d_{mn} = min_{i,j} d(M_i, M_j)$, distancia entre las dos medias más cercanas
M_q, media del conglomerado más cercano a x_j (condición 2)
M_p, media del segundo conglomerado más cercano a x_j (condición 2)

{Condición 1}
1: $centroideReemplazado \leftarrow$ **false**
2: **if** $min_i d(x_j, M_i) > d_{mn}$ **and** $d(x_j, M_m) > d(x_j, M_n)$ **then**
3: $M_n \leftarrow x_j$
4: $centroideReemplazado \leftarrow$ **true**
5: **end**
6: **if** $min_i d(x_j, M_i) > d_{mn}$ **and** $d(x_j, M_m) < d(x_j, M_n)$ **then**
7: $M_m \leftarrow x_j$
8: $centroideReemplazado \leftarrow$ **true**
9: **end**
{Condición 2}
10: **if** $centroideReemplazado =$ **true then**
11: **if** $d(x_j, M_p) > min_i d(M_q, M_i)$ **then**
12: $M_q \leftarrow x_j$
13: **end**
14: **end**

para maximizar la calidad de los resultados (ver paso de preparación de los datos en sección 3.3.2).

Además, también es de vital importancia identificar un conjunto de variables adecuada para el análisis de conglomerados, es decir, un conjunto de características a partir de los cuales los perfiles de usuario puedan ser diferenciados entre sí. Esta tarea se encuentra íntimamente relacionada con la descrita en la sección 3.3.2 Preparación de los datos, donde se reduce la dimensionalidad de los datos para disminuir la complejidad asociada a las técnicas de minería de datos y aumentar la calidad de los resultados. Esta tarea ha sido realizada con la ayuda del análisis de frecuencias y de correlaciones descritos en la sección 3.4.2.

En la figura 3.7 se muestra el proceso iterativo que se ha realizado y que cuenta con dos bucles de realimentación, a partir de los cuales se seleccionan distintos conjunto de variables y se ajustan los parámetros de entrada del algoritmo *K-Means* (selección de K óptimo). Debido a la existencia de múltiples soluciones sub-óptimas, cada ejecución del algoritmo con diferentes parámetros puede proporcionar resultados que no tienen por qué ser necesariamente incorrectos [Peña et al., 1999]. Por esta razón, se requiere realizar varias iteraciones entre este paso y el siguiente paso del KDP (sección 3.5 Evaluación del conocimiento extraído). Este proceso iterativo facilita la identificación

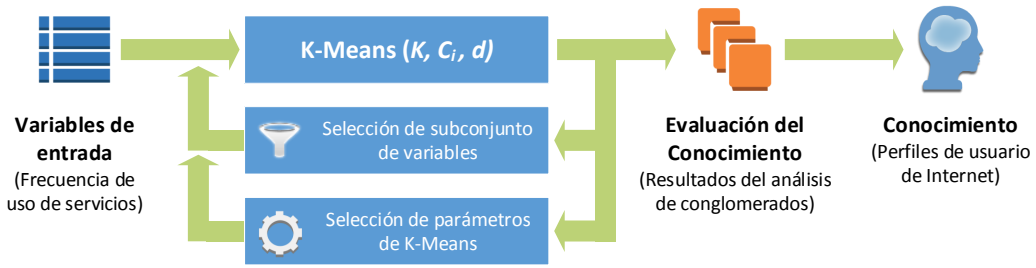

Figura 3.7: Proceso iterativo de análisis de conglomerados para identificar la tipología de usuarios de Internet

de una solución de alta calidad desde el punto de vista objetivo (mediante el uso de métricas de calidad objetivas) y desde el punto de vista conceptual (comparándolo con los resultados obtenidos en otros estudios de la literatura).

Para el caso de estudio de los usuarios residenciales en España, se aplica el proceso descrito anteriormente con el objetivo de obtener una solución de alta calidad para un subconjunto determinado de variables y un número óptimo de conglomerados K. Esta solución proporciona los conglomerados descubiertos en los datos, es decir, un conjunto de perfiles existentes entre los usuarios residenciales en España.

Los parámetros de entrada utilizados para la solución de mayor calidad encontrada son los siguientes:

- Número de conglomerados: $K = 4$
- Número máximo de iteraciones hasta convergencia de centroides: valor sufiencentemente alto para asegurar que el algoritmo converja a una solución
- Número de muestras utilizadas como datos de entrada: $N = 77896$
- Variables de frecuencia de uso de servicios en Internet de tipo escalar (para adecuarse al algoritmo):
  - Correo electrónico
  - Compartición de archivos
  - Juegos en red
  - Redes sociales
  - Lectura de información
  - Programas y series de TV

Variables	Conglomerados			
Uso semanal de servicios (días/semana)	1	2	3	4
Correo electrónico	2,55	4,91	4,78	5,19
Compartición de archivos	0,22	0,46	0,65	1,03
Juegos en red	0,20	0,24	0,65	0,78
Redes sociales	0,43	0,38	5,24	5,46
Lectura de información	0,49	4,88	0,57	5,43
Programas y series de TV	0,24	0,61	0,87	1,74
Operaciones con entidad bancaria	0,25	0,87	0,36	1,05
Compra de productos y servicios	0,06	0,20	0,12	0,34
Porcentaje de casos / conglomerado (%)	31 %	18 %	27 %	24 %

Tabla 3.5: Valores medios de frecuencia de uso de servicios para cada conglomerado

- Operaciones con entidad bancaria
- Compra de productos y servicios

La tabla 3.5 muestra los valores medios de las variables utilizadas como parámetros de entrada en el algoritmo y con un valor de K igual a 4. La unidad de medida representada en la tabla es el número de días en el que un servicio ha sido utilizado a lo largo de una semana. El segmento de *No-Usuarios* no ha sido incluido en la tabla debido a que fue identificado en la sección anterior.

Los resultados de este paso del KDP se describen en detalle en la sección 3.6 Resultados: conocimiento descubierto. Esta sección también interpreta los perfiles de usuario extraídos basándose en los patrones de uso de servicios de Internet.

3.5. Evaluación del conocimiento extraído

En esta sección se evalúan los resultados obtenidos a partir de la aplicación de técnicas de minería de datos mediante métricas de calidad objetivas. Además, también se realiza un análisis comparativo de los resultados frente a otros estudios disponibles en la literatura.

3.5.1. Análisis de Métricas de Calidad

Las métricas de calidad objetivas utilizadas analizan las características intrínsecas compartidas, o no, por los diferentes conglomerados extraídos de las técnicas de minería de datos. Existen dos criterios que son utilizados habitualmente por las métricas de evaluación más relevantes y populares de la literatura: la cohesión y separación de conglomerados. Además de estos dos criterios, es importante que la métrica de calidad

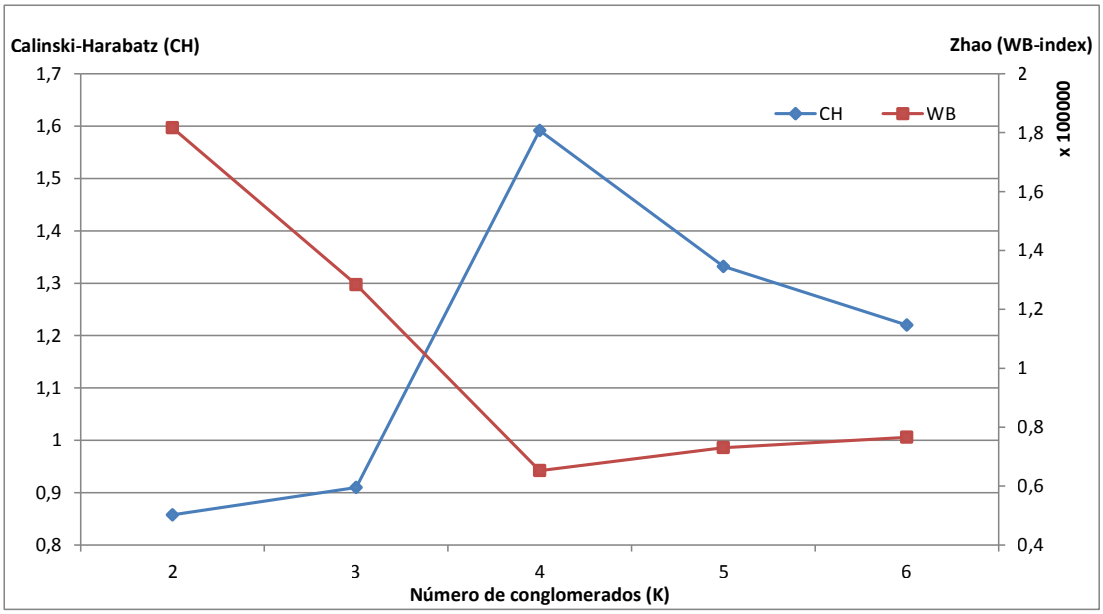

Figura 3.8: Aplicación de métricas para la evaluación de calidad de conglomerados

también tenga en cuenta el número de conglomerados extraídos, es decir, aquellos modelos con un número excesivo de perfiles de usuarios deberían de ser penalizados.

Como se describe en la sección 2.1.7, en la literatura existen multitud de métricas para la evaluación de la calidad de un análisis de conglomerados [Milligan and Cooper, 1985]. En esta tesis doctoral se selecciona la métrica de Caliński-Harabasz (CH) [Caliński and Harabasz, 1974] debido a su baja complejidad computacional y los buenos resultados obtenidos en diferentes estudios comparativos [Baarsch and Celebi, 2012, Tibshirani et al., 2001, Milligan and Cooper, 1985]. Además, también se utiliza una métrica similar llamada índice-WB, definida en [Zhao et al., 2009]. La calidad del análisis de conglomerados es directamente proporcional a la métrica CH e inversamente propocional al índice-WB.

La aplicación de estas dos métricas a los resultados de varias iteraciones de la fase de minería de datos se muestra en la figura 3.8. Cuando el número de conglomerados es igual a 4, es decir, $K = 4$ (el segmento correspondiente a los *No-usuarios* no se incluye), la métrica CH alcanza su máximo y el índice-WB alcanza su mínimo. Esto indica que según estas dos métricas, existe un máximo de calidad para este número de segmentos.

3.5.2. Análisis comparativo con otros estudios

Los resultados obtenidos en la fase de minería de datos del KDP también pueden ser evaluados a partir de un análisis comparativo con otros estudios similares y disponibles

en la literatura. En contraste con el enfoque cuantitativo del análisis de calidad mediante métricas de calidad objetivas, esta evaluación comparativa es de naturaleza cualitativa ya que presta especial atención a las diferencias y similitudes entre perfiles extraídos y los presentes en otros estudios.

Los objetivos de esta evaluación comparativa son los siguientes:

1. Confirmar la coherencia y validez de la tipología de usuarios de Internet extraída
2. Remarcar las nuevas contribuciones que han sido encontradas en el conocimiento extraído del KDP

Debido a que para el análisis comparativo con otros estudios se requiere conocer las características intrínsecas de los diferentes perfiles de usuario de Internet, esta evaluación se presenta conjuntamente en la sección 3.7 Discusión: uso del conocimiento extraído.

3.6. Resultados: conocimiento descubierto

Esta sección presenta los resultados derivados de la aplicación del proceso iterativo de análisis de conglomerados para identificar los perfiles de usuarios de Internet residenciales en España. En primer lugar, se describe el segmento de *No-Usuarios* comparándolo con el segmento correspondiente a los usuarios de Internet. Posteriormente, se desglosa este segundo segmento con los resultados obtenidos del análisis de conglomerados utilizando el algoritmo *K-Means*, identificando y describiendo los perfiles de usuarios de Internet extraídos. Por último, la sección concluye con una caracterización socio-demográfica de la tipología de usuarios de Internet.

3.6.1. Tipología de usuarios de Internet

En la sección 3.4.1, los *No-Usuarios* se identifican a partir de la extracción de información directamente de las variables de los datos de entrada. Un subconjunto de variables se utilizó para filtrar aquellos usuarios con acceso de banda ancha en sus hogares y su frecuencia de uso de Internet. A partir de esta identificación, se puede derivar el segmento complementario, el cual representa a los usuarios de Internet.

- No-Usuarios (47 %): esta categoría incluye a los individuos que no usan Internet de forma regular (el umbral temporal considera es un año sin acceder a Internet) o no tienen acceso en sus hogares. Este grupo de individuos se puede clasificar a su vez en los siguientes tipos:
  - Individuos que no usan Internet y sin acceso en el hogar (31,7 %).
  - Individuos que no usan Internet y con acceso en el hogar (7,9 %).

- Individuos que sí utilizan Internet pero no tienen acceso en el hogar (7,2 %).

- Usuarios de Internet (53 %): esta categoría incluye a individuos que usan Internet desde sus hogares de forma regular.

Los resultados del análisis de conglomerados, descritos en la sección 3.4.3, identifican 4 perfiles de comportamiento de usuarios de Internet. Cada conglomerado es analizado y etiquetado conforme a sus principales características relativas a los patrones de uso de servicios de Internet. Como se ilustra en la figura 3.9, la tipología de usuarios de Internet, incluyendo a los *No-Usuarios*, cumple con los objetivos del KDP y de minería de datos, definidos en la sección 3.2.3. A continuación, se etiquetan y describen los conglomerados correspondientes a los Usuarios de Internet:

- Conglomerado 1 - Usuarios Esporádicos (31 %, 16 % del total): este tipo de usuarios se encuentran caracterizado por un uso ocasional e infrecuente de servicios de Internet, normalmente servicios orientados a la comunicación, como por ejemplo, correo electrónico, mensajería instantánea, etc. Casi la mitad de ellos utiliza Internet de forma diaria.

- Conglomerado 2 - Usuarios Instrumentales (18 %, 10 % del total): estos usuarios se caracterizan por tener un uso muy intensivo de servicios orientados a objetivos, como por ejemplo, búsqueda de información, banca electrónica y compras online. Más del 90 % de ellos usan Internet de forma diaria.

- Conglomerado 3 - Usuarios Sociales (27 %, 14 % del total): estos usuarios se caracterizan por un uso muy intensivo de servicios de redes sociales, en comparación con el uso que le dan a otros servicios de Internet. Además, este perfil de usuario también tiene un uso por encima de la media de servicios de entretenimiento, como por ejemplo, visionado de series y televisión, y juegos en red. Casi el 95 % de estos usuarios se conectan de forma diaria a Internet.

- Conglomerado 4 - Usuarios Avanzados (24 %, 13 % del total): este tipo de usuarios cuentan con los valores de frecuencia de uso más alto en todos los servicios. Son usuarios con un patrón de consumo de servicios de Internet variado e intenso. Son los más usan los servicios más específicos y complejos, como por ejemplo, compartición de archivos, visionado de series y televisión, etc. Más del 98 % de ellos usan Internet de forma diaria.

La figura 3.10 muestra la frecuencia de uso medio semanal para todos los servicios de Internet para los 4 perfiles de usuario de Internet, incluyendo aquellas variables (servicios) que no fueron utilizadas como parámetros de entrada para el algoritmo de minería de datos. Este es el caso de los siguientes servicios: mensajería instantánea, telefonía sobre Internet, blogs y foros, escuchar música y podcasts.

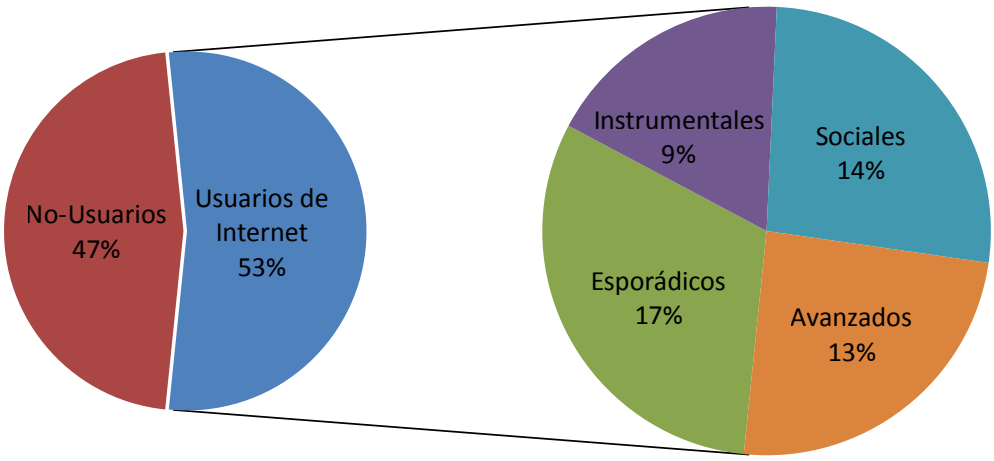

Figura 3.9: Tipología de usuarios de Internet residenciales en España

3.6.2. Caracterización demográfica de los usuarios de Internet

Como parte del conocimiento descubierto a partir del KDP, se realiza una caracterización de los conglomerados identificados en términos de variables socio-demográficas. A parte de los patrones de comportamiento, los perfiles de usuario también difieren en características socio-demográficas.

La tabla 3.6 presenta las distribuciones de algunas de las variables socio-demográficas más relevantes para cada conglomerado de la tipología de usuario de Internet. Las características más relevantes de esta caracterización se detallan en las siguientes líneas.

No-Usuarios. Este es el segmento de mayor edad identificado en el KDP. La media de edad es de 55 años, 11 años más que la media de edad de los usuarios de Internet. Como una posible consecuencia de esta avanzada edad, el 50 % de estas personas no se encuentran empleados ni buscan trabajo y sólo el 36 % tienen trabajo. Otras variables características, como el nivel de educación o el hábitat son más bajos para los *No-Usuarios* que para los *Usuarios de Internet*. Posiblemente esto se deba a que estos individuos vivan en áreas menos pobladas y con menos educación. El alto número de hogares con un único miembro también llama la atención, siendo un 10 % más alto que comparado con los usuarios de Internet. Finalmente, destaca la uniformidad de género en este segmento.

Usuarios de Internet. Este conglomerado está formado por los diferentes perfiles de usuarios de Internet identificados a lo largo del KDP, es decir, está compuesto por los usuarios *Esporádicos*, *Instrumentales*, *Sociales* y *Avanzados*. Este segmento de usuarios es especialmente útil para caracterizar las características generales de los usuarios de Internet como una única entidad. La proporción entre mujeres (49 %) y hombres

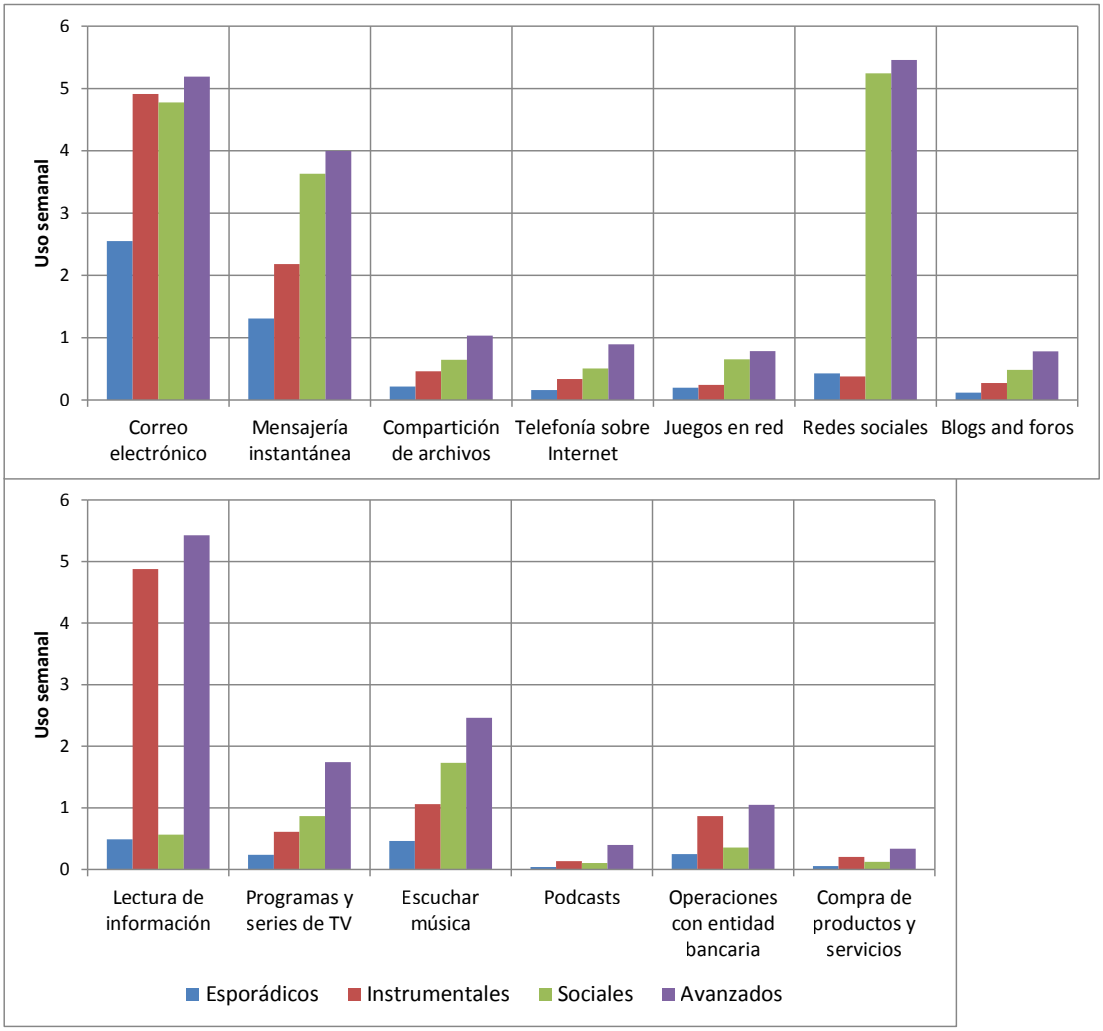

Figura 3.10: Uso medio semanal (días por semana) de los perfiles de usuario de Internet

Variables	Conglomerados					
		Usuarios de Internet				
	No-Usuarios	(Todos)	Esporádicos	Intru-mentales	Sociales	Avanza-dos
Género						
Hombre	43 %	51 %	45 %	62 %	44 %	56 %
Mujer	57 %	49 %	55 %	38 %	56 %	44 %
Edad						
14-19	3 %	8 %	5 %	4 %	14 %	9 %
20-24	3 %	7 %	4 %	4 %	11 %	10 %
25-34	11 %	22 %	17 %	16 %	25 %	28 %
35-44	14 %	26 %	27 %	25 %	24 %	27 %
45-54	15 %	17 %	22 %	24 %	11 %	12 %
55-64	17 %	10 %	13 %	15 %	6 %	6 %
65+	37 %	10 %	12 %	13 %	7 %	6 %
Miembros en la familia						
1	20 %	9 %	10 %	10 %	8 %	10 %
2	33 %	23 %	23 %	28 %	21 %	23 %
3	22 %	29 %	30 %	28 %	30 %	28 %
4+	25 %	38 %	37 %	34 %	41 %	39 %
Habitat						
Hasta 2k	9 %	5 %	6 %	4 %	5 %	4 %
De 2k a 5k	9 %	6 %	6 %	5 %	5 %	5 %
De 5k a 10k	9 %	7 %	9 %	6 %	7 %	6 %
De 10k a 50k	26 %	22 %	25 %	21 %	23 %	18 %
De 50k a 200k	22 %	26 %	25 %	27 %	25 %	27 %
De 200k a 500k	12 %	15 %	14 %	17 %	14 %	14 %
Más de 500k	12 %	20 %	15 %	18 %	21 %	25 %
Nivel de educación						
Sin estudios	9 %	1 %	1 %	1 %	2 %	1 %
Cert. escolar (10)	13 %	3 %	4 %	4 %	3 %	2 %
2º Grado (18)	66 %	69 %	74 %	63 %	70 %	64
Ed. Superior	12 %	27 %	21 %	32 %	25 %	33 %
Estado laboral						
Empleado	36 %	55 %	55 %	59 %	50 %	59 %
Desempleado	14 %	18 %	17 %	15 %	21 %	19 %
No trabaja/busca	50 %	27 %	28 %	26 %	29 %	23 %

Tabla 3.6: Variables socio-demográficas de la tipología de usuarios de Internet

(51 %) es prácticamente igual y la edad media es de 41 años. Como se ha mencionado anteriormente, los *Usuarios de Internet* tienden a vivir en zonas más pobladas que los *No-Usuarios*. Además, seguramente debido a la menor edad media, las tasas de empleo son mayores para éstos usuarios (55 %) que para los *No-Usuarios* (36 %). Finalmente, en términos de nivel de educación, el 27 % de los usuarios de Internet tienen una mayor educación en contraste con el 12 % de los que no utilizan Internet desde sus hogares.

Usuarios Esporádicos. Este conglomerado de usuarios tiene la peculiar característica de que la proporción de mujeres es ligeramente superior (55 %) al número de hombres. En términos de edad media, este segmento se encuentra ligeramente por encima de la edad media de los usuarios de Internet. El nivel de educación es más bajo que la media. Sólo el 21 % se los *Usuarios Esporádicos* tiene una educación superior, mientras que el 74 % disponen de segundo grado. Además, este conjunto de usuarios tiene un bajo nivel de hábitat, es decir, tienen a vivir en áreas poco pobladas. El resto de variables (miembros en el hogar y situación laboral) no son especialmente distintivas en este segmento.

Usuarios Instrumentales. Este perfil de usuario de Internet tiene la mayor diferencia de proporción de hombres y mujeres, con un 62 % de hombres frente al 38 % de mujeres. Estos usuarios son los que mayor edad media tiene (46 años) y el menor tamaño de los hogares (3 miembros por hogar). En términos de nivel de hábitat, los *Usuarios Instrumentales* se encuentran ligeramente por encima que los *Usuarios Esporádicos*, ya que tienen mayor tendencia a vivir en zonas con más población. En cuanto al nivel de educación, estos usuarios son el segundo segmento con mayor nivel (32 % educación superior y 63 % educación de segundo grado). Este segmento también destaca por tener la tasa de empleo más alta de todos los perfiles de usuarios de Internet.

Usuarios Sociales. Estos usuarios tienen la media de edad más baja, con tan sólo 36 años. Este segmento se compone de un 56 % de mujeres y un 44 % de hombres. También es característico de los *Usuarios Sociales* que suelen vivir en hogares con más miembros (3.2 miembros por hogar de media). Adicionalmente, este segmento cuentan con un nivel de hábitat algo superior al de los *Usuarios Instrumentales*, pero con peor nivel de educación y peor situación laboral.

Usuarios Avanzados. Estos usuarios constituyen el segundo segmento más joven de los usuarios de Internet, con una media de 37 años. En términos de distribución de género, hay un 56 % de hombres frente a un 44 % de mujeres. Este perfil de usuario de Internet tiene los valores más altos de nivel de educación (64 % de educación superior y 33 % de educación de segundo grado). También destaca este segmento tiene tendencia

a residir en zonas con mayor densidad demográfica. El tamaño de los hogares es alto con una media de 3.1 miembros. De forma similar a los *Usuarios Instrumentales*, la situación laboral de los *Usuarios Avanzados* es relativamente buena, con un 59 % de usuarios empleados.

3.6.3. Tecnología de acceso y caracterización de conexiones a Internet

Con el fin de complementar el conocimiento descubierto por el KDP, a continuación se provee de una visión de cómo los diferentes perfiles de usuario acceden a Internet. La tabla 3.7 presenta algunas de las variables de la colección de datos que proporcionan información acerca de la tecnología de acceso y las características de las sesiones de Internet de los usuarios. Más concretamente, se muestran las distribuciones de las variables de las tecnologías de acceso que utilizan los usuarios para conectarse a Internet. Además, en la tabla también se presenta la localización física desde la cual acceden a Internet y la duración media de la última sesión. La duración de la última sesión proviene de una cuestión dónde se preguntaba a los usuarios sobre la duración de la sesión en el día de ayer.

Variables	Conglomerados: Usuarios de Internet				
	(Todos)	Esporádicos	Intrumentales	Sociales	Avanzados
Frecuencia de acceso a Intenet					
Diaria	82 %	50 %	94 %	95 %	98 %
Varias veces por semana	13 %	32 %	5 %	5 %	1 %
Con menor frecuencia	6 %	18 %	0 %	0 %	0 %
Tecnología de acceso a Intenet					
Sin acceso	0 %	0 %	0 %	0 %	0 %
Esporádico (WiFi+Móvil)	9 %	10 %	5 %	12 %	7 %
xDSL	79 %	80 %	82 %	76 %	77 %
Cable	12 %	10 %	12 %	11 %	15 %
Otros	1 %	1 %	1 %	1 %	1 %
Localización de conexión					
Hogar	78 %	51 %	88 %	90 %	91 %
Trabajo	12 %	7 %	22 %	7 %	49 %
Centro de estudios	1 %	0 %	1 %	1 %	5 %
Otros sitios	8 %	3 %	5 %	14 %	33 %
Tiempo medio de duración de conexión desde localización (minutos)					
Hogar	127	88	109	131	164
Trabajo	191	153	192	143	233
Centro de estudios	108	117	136	73	121
Otros sitios	103	124	90	85	126

Tabla 3.7: Variables sobre acceso a la red de la tipología de usuarios de Internet

Frecuencia de acceso a Internet. Los *Usuarios Esporádicos* son los que se conectan con menor frecuencia a la red, con un 50 % de forma diaria frente al 82 % de media de los usuarios de Internet. El resto de perfiles se conectan de forma diaria con una proporción de usuarios cercana al 95 %, siendo los *Usuarios Avanzados* el conglomerado que accede a Internet con mayor asiduidad.

Tecnología de acceso a Internet. Como se puede apreciar, la tecnología predominante de acceso a Internet es xDSL. Casi el 80 % de los usuarios de Internet utilizan esta tecnología para acceder a la red. La supremacía de xDSL se encuentra ligeramente atenuada en los *Usuarios Sociales* debido al uso que hacen de acceso esporádicos inalámbricos, como por ejemplo puntos de acceso WiFi o la red celular móvil. De forma análoga, la proporción de uso de tecnología xDSL por parte de los *Usuarios Avanzados* es algo menor comparado con la totalidad de los usuarios de Internet, debido a que muchos de éstos acceden a través de redes de cable.

Localización de conexión. Los *Usuarios Esporádicos* son los que menos acceden a Internet en términos generales. En la tabla se aprecia como sólo el 51 % y el 7 % de estos usuarios acceden a la red desde sus hogares y desde el trabajo respectivamente. El resto de perfiles de usuarios de Internet acceden a Internet en una proporción cercana al 90 %. Más concretamente, los *Usuarios Avanzados* tienen la proporción más alta de conexión, desde el hogar con un 91 % y desde el trabajo con un 49 %. Por el contrario, sólo el 22 % de los *Usuarios Instrumentales* y el 7 % de los *Usuarios Sociales* y *Esporádicos* acceden a la red desde el trabajo.

Duración de la conexión. La duración de la conexión varía entre los diferentes perfiles de Internet y se ve íntimamente influenciada por la localización desde la cual se realiza el acceso. Destaca que todos los usuarios de Internet se conectan por más tiempo desde el trabajo que desde el hogar. Los *Usuarios Avanzados* cuentan con los tiempos de duración más altos, independientemente de la localización del acceso. Los *Usuarios Sociales* es el segmento segmento con mayores duraciones de conexión desde el hogar. A pesar de que los *Usuarios Instrumentales* se encuentran por detrás de los *Usuarios Sociales* en tiempo de conexión desde el hogar, éstos tienen tiempos de conexión mayores desde otras localizaciones, como por ejemplo, desde el trabajo o centro de estudios. Los *Usuarios Esporádicos* destacan por disponer de los menores tiempos de conexión medios desde el hogar, a pesar de que sus tiempos desde otras localizaciones son bastante altos.

3.7. Discusión: uso del conocimiento extraído

En esta sección se aborda el último paso de la metodología propuesta durante este capítulo de la tesis doctoral y cuyo objetivo es la utilización del conocimiento extraído durante el KDP. Gracias a los pasos previos se ha extraído un conjunto de perfiles de usuarios de Internet que representan los diferentes patrones de comportamiento y consumo de servicios para distintos grupos socio-demográficos.

Centrándose en el estudio llevado a cabo en esta tesis doctoral de los usuarios de Internet residenciales en España, el consecuente análisis de perfiles de usuarios conduce a algunas observaciones interesantes. Por ejemplo, destaca que el 47 % de la población no utiliza Internet desde sus hogares, es decir, que apenas la mitad de la población puede considerarse que sean usuarios residenciales de Internet. Este hecho permite afirmar que las TIC, y en especial Internet, todavía no cuentan con una penetración total en las vidas de los españoles. El porcentaje de *No-Usuarios* es ligeramente superior a las cifras reportadas en estudios previos, los cuales estiman que se encuentra entre el 42 % [Brandtzæg et al., 2011] y el 44 % [Ortega Egea et al., 2006]. En cualquier caso, estas cifras han de ser consideradas con cautela ya que los datos estadísticos utilizados en estos estudios se corresponden a los años 2006 y 2002 respectivamente.

Continuando el análisis, el segundo segmento que más predomina es el de los *Usuarios Esporádicos*, que representa un 16 % de la población. Este perfil de usuario accede a Internet de forma ocasional para utilizar servicios básicos, principalmente orientados a la comunicación. Este segmento de usuarios ya había sido identificado hace años en otros estudios [Selwyn et al., 2005]. La proporción de individuos en este conglomerado se encuentra en línea con los porcentajes presentes en otros estudios recientes en Europa. Por ejemplo, en [Brandtzæg et al., 2011] se describe a este segmento con una proporción del 18 %, mientras que en [Ortega Egea et al., 2006] se identifica un segmento similar, llamado *Rezagados*, con un tamaño del 16 % de la población.

El siguiente perfil de usuario más relevante es el de los *Usuarios Instrumentales*, que cuenta con el 10 % de la población de estudio. Existe un número significativo de personas que utilizan Internet con asiduidad y con unos fines muy específicos. Este grupo se caracteriza por acceder a la red principalmente desde sus hogares. Este patrón de comportamiento ya fue identificado por otros estudios como por ejemplo en [Howard et al., 2001] o [Brandtzæg, 2010]. A pesar de que utilizan la red con cierta frecuencia, Internet aún no se ha convertido en una parte esencial de sus vidas.

El conglomerado de *Usuarios Avanzados*, formado por el 13 % de la población, es uno de los segmentos más mencionados en otros estudios de la literatura. [Howard et al., 2001] ya identifica a este segmento, etiquetado como *Internautas*, y lo define como un grupo de usuarios con experiencia y que utilizan a diario muchos servicios de Internet de forma intensa. El porcentaje de este tipo de usuarios se encuentra en línea

con los tamaños de segmentos similares de otros estudios, como por ejemplo, el 12 % en [Brandtzæg et al., 2011] y el 19 % en [Ortega Egea et al., 2006].

Los *Usuarios Sociales*, con un 14 % de la población, es probablemente el descubrimiento principal de los resultados del análisis de conglomerados realizado durante el KDP. No se han encontrado referencias en la literatura que reporten la existencia de este perfil de usuario de Internet. No obstante, sí que existen estudios que analizan qué tipos de usuarios existen dentro de las redes sociales y comunidades online [OFCOM, 2008, Brandtzæg, 2010]. Es muy probable que la razón detrás de este fenómeno se encuentre en que las redes sociales se han establecido como una parte fundamental de la vida de muchos usuarios de Internet, lo cual ha provocado que exista un grupo de personas con unos patrones de uso de Internet muy característicos.

Por último cabe mencionar que en esta tesis doctoral no se ha identificado ningún grupo de usuarios, cuya motivación principal sea la del entretenimiento. A pesar de que los usos y gratificaciones asociados al entretenimiento han sido considerados en el análisis de conglomerados, parece que en España aún no existe un uso lo suficientemente generalizado de estos servicios como para que emerja un segmento significativo en la tipología de usuarios de Internet identificada durante el KDP. Sin embargo, debido al constante aumento del consumo de servicios de streaming de video, es muy probable que en un futuro cercano este perfil de usuario cobre una gran importancia y se incluya en la tipología de usuarios de Internet. Este afirmación se fundamenta en los pronósticos realizados por Cisco en [Cisco, 2013, Cisco, 2014], donde se prevé que en 2017 casi el 75 % del tráfico mundial provenga de este tipo de servicios.

3.7.1. Evolución de la tipología de usuario de Internet

A continuación se analiza la evolución de la tipología de usuarios de Internet extraída con el objetivo de analizar la evolución que han experimentado los usuarios de Internet residenciales en España en los años recientes.

Para esta labor, se ha aplicado la misma metodología descrita a lo largo de todo el capítulo basada en KDPs pero con colecciones de datos correspondiente a los años 2010 y 2011. La principal particularidad en los KDPs reside en que los análisis de conglomerados son de tipo semi-supervisado, ya que el objetivo es encontrar perfiles de usuarios que compartan características con los identificados para el año 2012. Gracias a esta particularidad, se puede realizar un análisis comparativo de la tipología de usuarios de Internet en los últimos años.

Los parámetros utilizados en la fase de minería de datos son idénticos a los utilizados anteriormente para la última anualidad. Sólo se diferencia en que se imponen unos centroides iniciales que coinciden con los centroides finales extraídos del análisis del año 2012. De esta forma, se fuerza que los segmentos identificados para años anteriores se

	2010	2011	2012
No-Usuarios	54 %	51 %	47 %
Usuarios de Internet	46 %	49 %	53 %
Esporádicos	20 %	19 %	16 %
Instrumentales	12 %	10 %	10 %
Sociales	7 %	11 %	14 %
Avanzados	7 %	9 %	13 %

Tabla 3.8: Evolución de tipología de usuario de Internet desde el 2010 al 2012

asemejen en características con los identificados en la sección 3.4.3.2.

La tabla 3.8 muestra las proporciones de los perfiles de usuario de Internet identificados en los años pasados. Se observa que la penetración de Internet es España ha ido aumentando paulatinamente a lo largo de los años. Este hecho se corrobora por el aumento en el número de usuarios de Internet a lo largo de los años analizados. Del mismo modo, esta tendencia puede observarse en las fluctuaciones porcentuales de los diferentes perfiles de usuario de Internet. Mientras que los segmentos con patrones de consumo de perfil bajo, como los *Usuarios Esporádicos* e *Instrumentales*, han disminuido a lo largo de estos años; los *Usuarios Avanzados* y *Sociales* han aumentado significativamente. La principal razón detrás de estos cambios en la tipología de usuarios de Internet se encuentra en que la tecnología se convierte, cada vez más, en un componente indispensable en la sociedad.

En cuanto a la evolución en el consumo de los servicios de Internet, en la figura 3.11 se muestra el uso semanal medio para cada servicio durante tres años. El eje vertical muestra el número de días en el que los usuarios consumen de media un servicio. Destaca el aumento de consumo en servicios como las redes sociales y mensajería instantánea. Esto confirma que las redes sociales y las comunidades online se están consolidando como servicios esenciales de Internet.

Por el contrario, el consumo de otros servicios ha disminuido significativamente en los últimos años, como es el caso de los servicios de compartición de archivos y visionado de videos en línea. Es muy probable que esta tendencia se encuentre relacionada con cambios en la legislación española asociados a los derechos de autor y que regula el consumo de este tipo de servicios en Internet.

3.7.2. Pronóstico de la tipología de usuario de Internet

El objetivo de esta sección consiste en realizar un pronóstico de la evolución de los perfiles de usuario de Internet en los próximos años en base a los resultados obtenidos a partir de la aplicación de varios KDPs para los últimos 3 años.

El pronóstico parte de la suposición que los perfiles de usuario no van a cambiar de

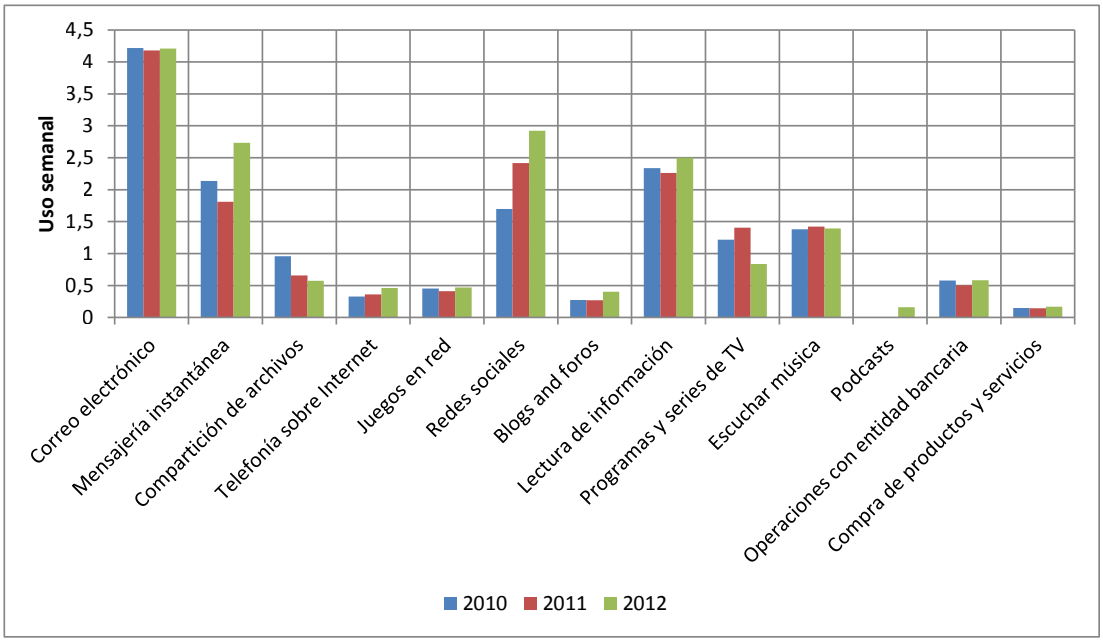

Figura 3.11: Uso medio semanal (días por semana) durante los últimos tres años de los perfiles de usuario de Internet

forma brusca en los próximos años y que las tasas de crecimiento de los servicios de Internet, positivas o negativas, tendrán cierta tendencia a suavizarse con los años. De acuerdo con estas premisas, se opta por realizar un ajuste de las curvas de evolución de los perfiles de usuario de Internet mediante una tendencia logarítmica (ecuación 3.2).

$$y(x) = \alpha \cdot ln(x) + \beta \tag{3.2}$$

En la tabla 3.9 se presentan los parámetros utilizados en la función de tendencia logarítmica (ecuación 3.2), indicando los parámetros de ajuste para cada tipo de perfil de usuario de Internet y el coeficiente de determinación R^2. Además, a modo ilustrativo, se indican las proporciones de los perfiles de usuario para el año 2017.

La figura 3.12 muestra la evolución pronosticada de los segmentos correspondientes a los *No-Usuarios* y *Usuarios de Internet*. Basado en los resultados de los KDPs entre los años 2010 y 2012, donde se aprecia un aumento de usuarios de Internet del 46 % al 53 %, se estima que exista un crecimiento gradual hasta el 2017 alcanzando un porcentaje cercano al 59 %. De forma complementaría, la proporción de *No-Usuarios* en la población descendería hasta el 41 %.

La figura 3.13 muestra la evolución prevista de los perfiles de usuario de Internet. De acuerdo con los pronósticos, se espera que en los siguientes 5 años, los *Usuarios*

	α	β	R^2	**2017**
No-Usuarios	-0,065	0,5476	0,9131	41 %
Usuarios de Internet	0,065	0,4524	0,9131	59 %
Esporádicos	-0,032	0,2045	0,9270	14 %
Instrumentales	-0,026	0,1225	0,9743	7 %
Sociales	0,0653	0,0675	0,9933	20 %
Avanzados	0,0581	0,058	0,8984	18 %

Tabla 3.9: Pronóstico de la tipología de usuario de Internet en el año 2017

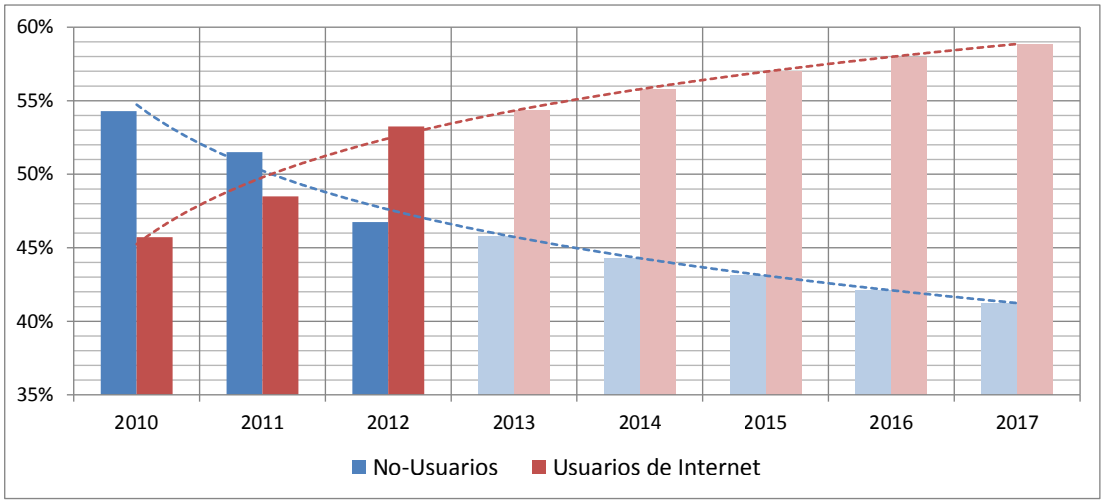

Figura 3.12: Pronóstico de No-Usuarios y Usuarios de Internet

Esporádicos e *Instrumentales* decrezcan hasta un 7 % y un 14 % respectivamente. Por el contrario, se espera un aumento de *Usuarios Sociales* y *Avanzados* hasta una proporción del 19 % y 17 % respectivamente. Las predicciones realizadas indican que en el año 2017, el segmento de usuarios más predominante será el correspondiente a los *Usuarios Sociales* seguido de cerca por los *Usuarios Avanzados*.

Es importante destacar el enfoque conservador utilizado en las predicciones anteriores. Se han realizado partiendo de la base de que en los próximos años no existan acontecimientos disruptivos que puedan causar cambios abruptos en los patrones de comportamiento de los usuarios de Internet, o que incluso puedan provocar la aparición de nuevos perfiles de usuarios. Por lo tanto, estas predicciones se han basado en que en los años venideros, las tasas de crecimiento siguen una tendencia moderada.

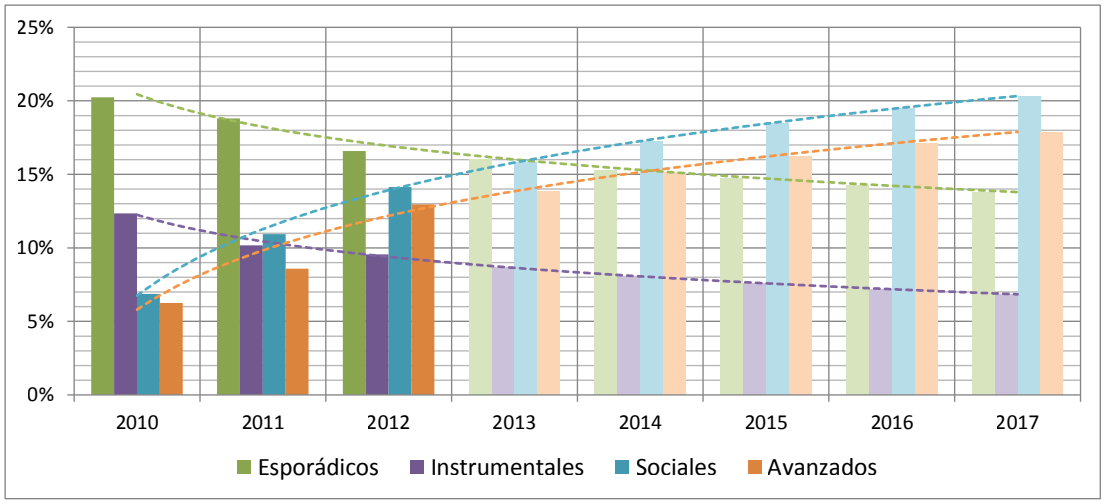

Figura 3.13: Pronóstico de perfiles de usuarios de Internet

3.8. Conclusiones

Existe un creciente interés en la identificación y comprensión de los patrones de consumo de los servicios de Internet y de la tipología de los usuarios asociados a los mismos. Esta información puede ser de gran utilidad para una amplia variedad de agentes de la industria de Internet, incluyendo operadores de red, proveedores de servicios o, incluso, empresas de publicidad.

En este capítulo se ha propuesto el uso de un KDP como marco metodológico robusto y conciso para la realización de una identificación y caracterización de los usuarios de Internet. Se han descrito y analizado todas las tareas de cada uno de los pasos secuenciales definidos en la metodología basada en KDP. La metodología aborda cómo realizar formalmente este tipo de estudios, desde la definición inicial del dominio del problema, seguido por la selección y el análisis de los datos de entrada, a la extracción, validación y aplicación de los resultados.

A lo largo del capítulo, se ha aplicado la anterior metodología al estudio de las tipologías de usuarios de Internet para el caso específico de usuarios residenciales en España. Esto proporciona un ejemplo práctico de cómo aplicar la metodología basada en KDP para la extracción de tipologías de usuarios de Internet.

Como principales conclusiones del capítulo, se ha identificado los perfiles de usuario de Internet existentes en diferentes colecciones de datos de entrada. Además, se han tenido en cuenta variables específicas al consumo de servicios en Internet, así como variables socio-demográficas para la posterior caracterización de los tipos de usuarios de Internet. La aplicación de KDP al conjunto de datos correspondientes al año 2012,

ha proporcionado la distinción entre los *Usuarios de Internet* (53 %) y *No-Usuarios* (47 %). Por otra parte, dentro del segmento de los usuarios de la red, el método es capaz de identificar cuatro grandes categorías: *Esporádicos* (16 %), *Instrumentales* (10 %), *Sociales* (14 %), y *Avanzados* (13 %).

Además, se ha aplicado la misma metodología para los datos estadísticos correspondientes a los últimos años, de forma que se ha realizado una predicción de la posible evolución de los perfiles de usuarios de Internet en los próximos años.

La aplicación de la metodología basada en KDP a la caracterización de los usuarios de Internet abre muchas líneas de posibles trabajos futuros. Es interesante considerar su aplicación al estudio de la conexión a Internet a través de dispositivos móviles. Desafortunadamente, a pesar del aumento en el uso de los teléfonos inteligentes y otros dispositivos móviles, todavía no existen datos estadísticos suficientemente completos para la realización de este tipo de estudios.

Los resultados de este capítulo serán utilizados como datos de entrada para el modelo de estimación de demanda de tráfico de Internet, descrito en los capítulos 4 y 5 de esta tesis doctoral.

Capítulo 4

Estimación de demanda de tráfico y dimensionado de red de acceso

4.1. Introducción

En este capítulo se describe la metodología utilizada para la estimación de demanda de tráfico y dimensionado de red de acceso, y los principales componentes en los que se descompone. La figura 4.1 muestra las entradas necesarias para aplicar la metodología, así como los principales modelos que se definen para la estimación de la demanda y el dimensionado:

- **Modelo de red de acceso:** en la sección 4.3.1 se describe como extraer un modelo genérico de red a partir de las arquitecturas de redes de acceso revisadas en la sección 2.3 Dimensionado de redes de acceso.

- **Modelo de tráfico de red:** en la sección 4.3.2 se desarrolla un modelo global de tráfico de red basado en la composición de modelos de tráfico de aplicaciones de Internet, descritos en la sección 2.2 Caracterización de tráfico de Internet.

- **Modelo de perfiles de usuario y aplicaciones:** en la sección 4.3.3 se desarrolla un modelo que caracteriza la demanda de uso de aplicaciones de usuarios de Internet a partir de los resultados extraídos en el capítulo 3.

Antes de describir con detalle todos los modelos definidos en la metodología, se describe el problema que se intenta abordar en este capítulo de la tesis doctoral: el acceso compartido a recurso de red. Posteriormente, se detalla cada modelo, y sus correspondientes parámetros de entrada, necesarios para la estimación de la demanda de tráfico en un escenario de red acceso.

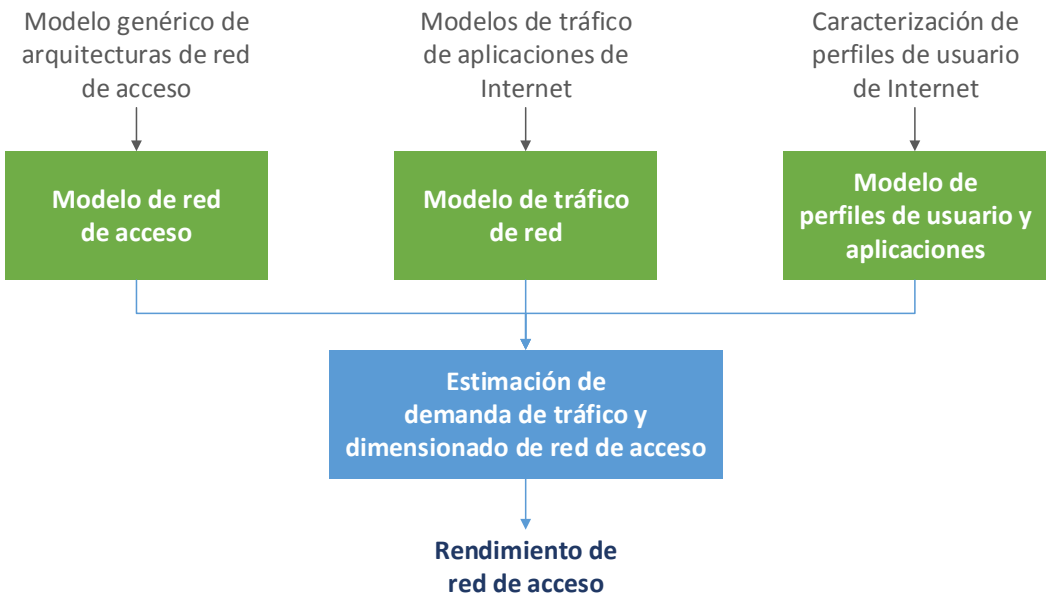

Figura 4.1: Entradas y salidas del método de estimación de demanda de tráfico y dimensionado de red de acceso

4.2. Acceso compartido a recurso de red

El tráfico de datos es inheretemente elástico debido a que su tasa de transferencia puede ser modulada en función de la demanda actual y, a menudo, sin afectar al rendimiento percibido por el usuario. En el caso de Internet, los protocolos a nivel de transporte, como por ejemplo TCP, tienen como objetivo explotar la capacidad total de una red que está siendo utilizada por varios usuarios. Por esta razón, nos encontramos ante un problema de compartición de ancho de banda.

4.2.1. Definición del problema

Del análisis realizado sobre las arquitecturas de diferentes tipos de redes de acceso (sección 2.3), se aprecia como existen varios niveles de agregación. Debido a este fenómeno, un número determinado de usuarios tienen que compartir recursos de red, como por ejemplo, los enlaces red y su correspondiente ancho de banda.

En la figura 4.2 se muestra un ejemplo en donde se considera un conjunto de suscriptores $H_1, H_2, ..., H_N$ conectados a una misma red de acceso. Estos usuarios han de compartir un mismo recurso de red, dado que su tráfico se agrega en un mismo enlace de la red de acceso (por ejemplo a través de un DSLAM que hace funciones de Multiplexor (MUX)).

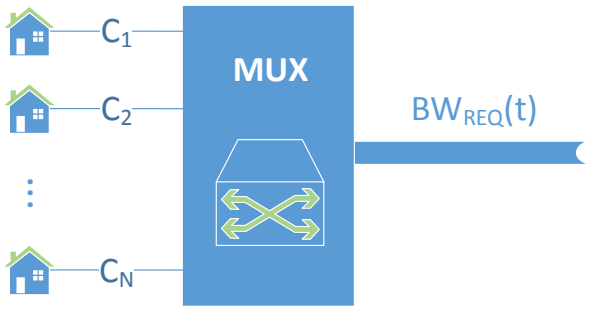

Figura 4.2: Esquema conceptual de compartición de ancho de banda entre N usuarios

Si suponemos que la velocidad percibida por los usuarios no se encuentra limitada por la red de acceso, se necesitará un ancho de banda en el enlace lo suficientemente grande como para soportar todo el tráfico agregado de los usuarios a lo largo del tiempo $BW_{REQ}(t)$. Así pues, si la red de acceso no limita la velocidad, la limitación de los suscriptores a la red, vendrá dado por la capacidad contrata con con sus proveedores de acceso a Internet, $C_1, C_2, ..., C_N$.

4.2.2. Superposición de actividad de suscriptores

Una vez descrito el problema de compartición de ancho de banda entre suscriptores, se describe con más detalle la superposición de actividad que determina el ancho de banda agregado necesario en el enlace compartido. Se modela la actividad de cada suscriptor de la red de acceso mediante periodos de actividad e inactividad, correspondientes a aquellos periodos en los que se transmitan datos o no.

En la figura 4.3 se muestra cómo se calcula el ancho de banda requerido necesario $BW_{REQ}(t)$ en el enlace. Se estima a partir de la superposición de los periodos de actividad de los suscriptores, cuyo ancho de banda de los suscriptores $H_1, H_2, ..., H_N$ se corresponde con su capacidad contratada $C_1, C_2, ..., C_N$.

En la ecuación 4.1 se define que en un tiempo cualquiera $t = t_i$, el valor del ancho de banda agregado requerido $BW_{REQ}(t = t_i)$ será igual a la suma de aquellos anchos de banda que se encuentren siendo utilizados por los suscriptores, que se corresponden a la capacidad contratada $C_1, C_2, ..., C_N$.

$$BW_{REQ}(t = t_i) = \alpha_1(t_i) \cdot C_1 + \alpha_2(t_i) \cdot C_2 + ... + \alpha_N(t_i) \cdot C_N : \alpha_j(t) \in \{0, 1\} \quad (4.1)$$

Debido a que los cambios en el ancho de banda agregado requerido sólo cambia cuando varía un periodo de actividad de algún suscriptor, BW_{REQ} puede describirse

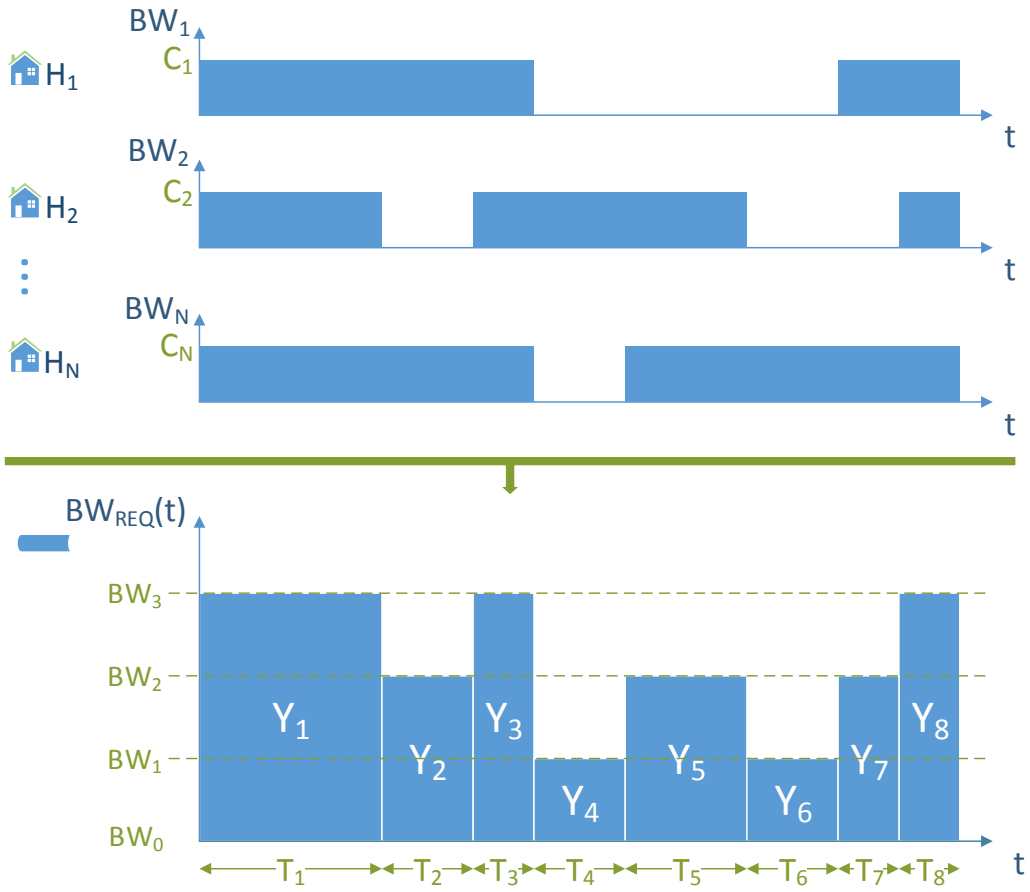

Figura 4.3: Superposición de actividad de N suscriptores y ancho de banda requerido

a partir de dos variables discretas: una que define el valor del ancho de banda y otra el periodo durante el cual se ha requerido ese ancho de banda. En la ecuación 4.2 se muestran los L valores que pueden tomar ambas variables discretas.

$$\begin{aligned} BW_{REQ} &= \{Y_1, Y_2, ..., Y_L\} \\ Timesteps &= \{T_1, T_2, ..., T_L\} \end{aligned} \tag{4.2}$$

4.2.3. Rendimiento de la red: Grado de Servicio (GoS)

En la sección 2.3 se puso de manifiesto que algunas métricas de rendimiento, como la velocidad o la capacidad, no eran suficientemente precisas para caracterizar el rendimiento percibido desde el punto de vista de los usuarios de la red.

En primer lugar, se parte de la hipótesis de que se ha analizado la actividad de N suscriptores durante un periodo de tiempo suficientemente largo para que sea representativo de la actividad habitual de los mismos. Bajo esta premisa, el término de GoS se define en función de los valores de ancho de banda agregado requerido en el enlace compartido BW_{REQ}.

Como se aprecia en la figura 4.3 los valores de BW_{REQ} son finitos ya que son una combinación lineal de las capacidades de los suscriptores. Por esta razón, los anchos de banda agregados requeridos podrían agregarse sin importar el orden de los periodos en los que fueron obtenidos, transformando las representaciones descritas en la ecuación 4.2 por las de la ecuación 4.3.

$$\begin{aligned} BW_{REQ} &= \{BW_0, BW_1, BW_2, ..., BW_M\} \\ T_{BW} &= \{T_{BW_0}, T_{BW_1}, T_{BW_2}, ..., T_{BW_M}\} \end{aligned} \tag{4.3}$$

En la figura 4.4 se muestra esta agregación de anchos de banda con sus correspondientes periodos. En este caso particular, el periodo del ancho de banda BW_0 es cero (T_{BW_0}) debido a que no ha existido ningún momento en el que todos los usuarios estuviesen inactivos.

Suponiendo que los valores de los anchos de banda $\{BW_0, BW_1, BW_2, ..., BW_M\}$ son crecientes, se estima el GoS correspondiente a un ancho de banda mediante la ecuación 4.4.

$$GoS(BW_i) = Prob(BW \geq BW_i) = \frac{1}{T_{Total}} \cdot \sum_{j=i}^{M} T_{BW_j} = \frac{\sum_{j=i}^{M} T_{BW_j}}{\sum_{j=0}^{M} T_{BW_j}} \tag{4.4}$$

Se define el GoS como la probabilidad en la que el ancho de banda en el enlace es mayor que cierto umbral BW_i. A modo de ejemplo, si se aplicase esta ecuación al ejemplo mostrado en la figura 4.4, se obtendría que $GoS(BW_0) = 1$ y $GoS(BW_1) = 1$,

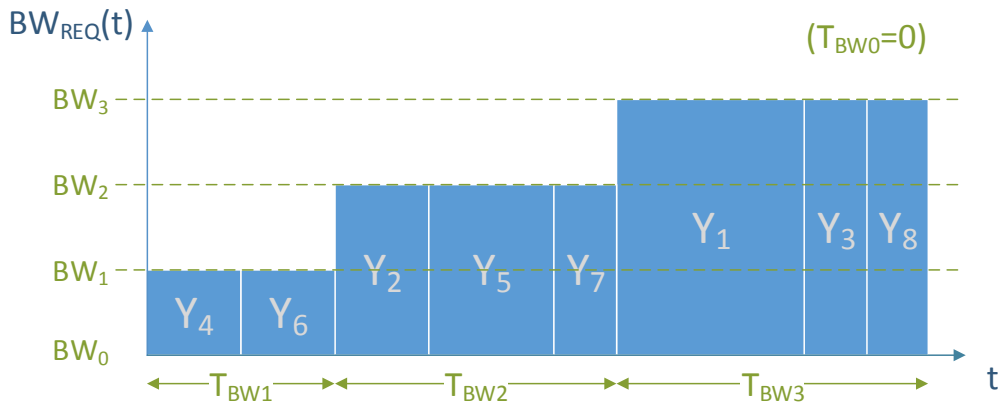

Figura 4.4: Ancho de banda requerido por N suscriptores ordenados y con periodo de tiempo asociado

ya que el ancho de banda agregado requerido en el enlace siempre ha sido igual o mayor a BW_1.

De forma análoga a la definición del GoS para los sistemas de telefonía, si un valor de ancho de banda BW_i produce un GoS de 0 %, implicaría que el ancho de banda agregado requerido por los usuarios nunca ha superado ese valor, es decir, los suscriptores siempre podrían utilizar su velocidad máxima. De forma contraria, si se obtuviese un GoS de 100 %, implicaría que el enlace de red estaría siempre saturado y los suscriptores no podrían utilizar su tasa de bit máxima.

4.2.4. Propuesta de modelo de estimación de rendimiento

A continuación se propone un modelo para estimar el rendimiento necesario en un enlace agregado de una red de acceso. En la figura 4.5 se ilustra un diagrama del modelo de cola que representa a los N suscriptores y al enlace que agrega todo el tráfico de los mismos. Los componentes del modelo son:

- N fuentes de tráfico de tipo ON/OFF heterogéneas que modelan la actividad de los suscriptores de las redes de acceso. Durante el periodo de actividad la tasa de bit disponible es de h_i (bps).
- 1 servidor con una cola de tamaño X que modela el enlace donde se agrega todo el tráfico de los suscriptores y que tiene una capacidad de BW_{REQ} (bps).

Las fuentes de tráfico de tipo ON/OFF se consideran heterogéneas debido a que un suscriptor de una red de acceso suele corresponderse a un hogar donde existen un

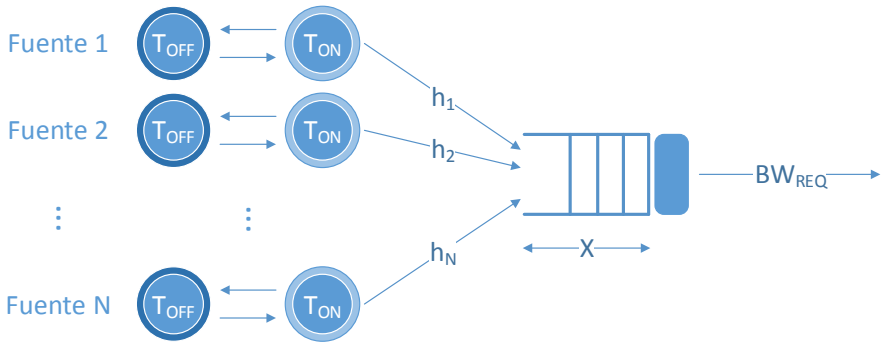

Figura 4.5: Esquema conceptual de compartición de ancho de banda de N suscriptores

conjunto de usuarios y que a su vez, pueden usar un conjunto de aplicaciones que hagan uso de la red. Además, como se verá en la siguiente sección, en algunos casos, un usuario puede hacer uso de varias aplicaciones simultáneas. Por estas razones, se considera que las fuentes de tráfico son diferentes entre sí.

El objetivo del modelo propuesto es la de caracterizar el enlace de tráfico agregado a partir del ancho de banda requerido BW_{REQ}, suponiendo que los suscriptores de la red utilizan la capacidad contratada. Por esta razón, las tasas de bit de los suscriptores h_i es igual a la capacidad contratada por cada uno de ellos C_i. Debido a que en este modelo, la capacidad del enlace agregado es variable, y a priori desconocida, se asume que la cola del servidor es infinita ($X \to \infty$) y que el canal siempre puede gestionar todo el tráfico agregado.

El objetivo del uso de este modelo es la de caracterizar el ancho de banda que sería necesario para dar servicio a los usuarios de una red de acceso que han contratado unas velocidades determinadas. Para ello se hace uso de la métrica de rendimiento anteriormente descrita, el GoS para diferentes umbrales de ancho de banda BW_i.

4.3. Metodología de estimación de demanda de tráfico

En esta sección presenta una metodología que permite realizar una estimación de la demanda de tráfico necesaria por un conjunto de usuarios y que tiene como principal objetivo dimensionar el ancho de banda necesario de una red de acceso.

4.3.1. Modelo de red de acceso

Un componente de la metodología de estimación de demanda de tráfico consiste en un modelo de red de acceso que permite describir el escenario de red bajo estudio.

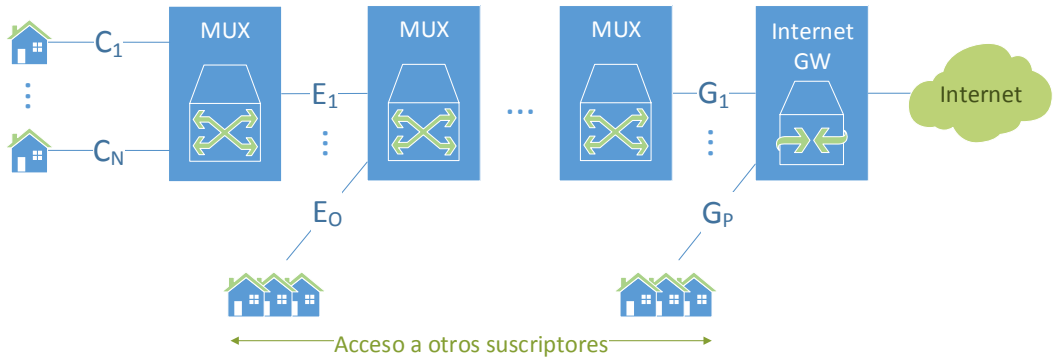

Figura 4.6: Esquema de red de acceso residencial con varios niveles de agregación

En la sección 2.3 se realiza una revisión del estado del arte de los conceptos necesarios para el dimensionado de redes de acceso, haciendo especial hincapié en las arquitecturas de red. Además, se define un modelo genérico de red (sección 2.3.3) válido para cualquier tecnología de red de acceso (figuras 2.19 y 2.20).

4.3.1.1. Selección de nivel de agregación

Recordando la definición del problema (sección 4.2.1), esta metodología busca analizar el ancho de banda requerido a lo largo del tiempo por un conjunto de usuarios que utilizan toda su capacidad de enlace $(C_1, C_2, ...C_N)$, así como el rendimiento de red percibido por los mismos en términos de GoS.

En la figura 4.6 se ilustra una red de acceso residencial que se compone de un conjunto finito de niveles de agregación y que presta servicio de conexión a un número determinado de suscriptores. Se puede apreciar como con cada nivel de agregación, aumenta el número de suscriptores soportados por los enlaces de la red.

En primer lugar, se ha destacar la dificultad asociada a la identificación de los cuellos de botella, ya que incluso puede darse el caso de que coexistan varios en función de los niveles de agregación de la arquitectura de red. Además, los cuellos de botellas no sólo dependen de la capacidad de los enlaces en la red de acceso, sino también de la demanda de tráfico de los usuarios a lo largo del tiempo.

Por este motivo, para el modelo de red de esta metodología se opta por una simplificación de la arquitectura de red de acceso, donde se estudia y analiza el rendimiento de un único enlace de agregación. En este modelo simplificado, se conoce el número de suscriptores a los que se da servicio y que esperan un rendimiento asociado a la capacidad que de acceso que tienen contratada $(C_1, C_2, ...C_N)$. Este esquema se corresponde al

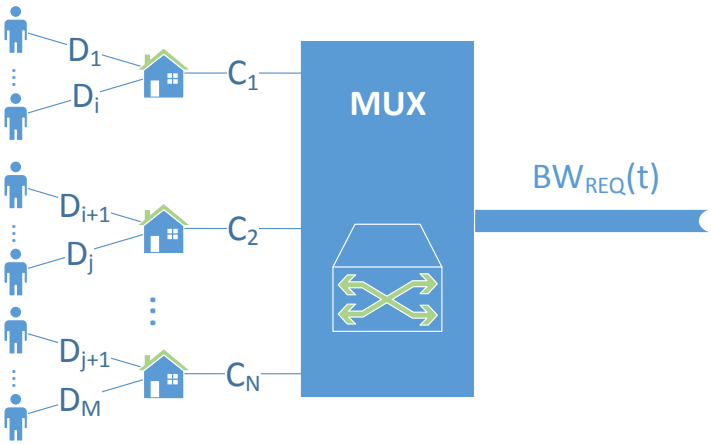

Figura 4.7: Diagrama de modelo de red de acceso con N suscriptores

anteriormente presentado en la figura 4.2.

En caso de que se utilizase esta metodología para dimensionar una red de acceso con diferentes niveles de agregación, se aplicaría de forma secuencial comenzando desde el menor nivel de agregación (última milla) hasta llegar hasta la pasarela de Internet (*Internet GW*). Para cada nivel de agregación se irían definiendo los anchos de banda requeridos (BW_{REQ}) para un nivel determinado de GoS.

4.3.1.2. Usuarios y suscriptores de red de acceso

Hasta el momento, se ha hecho alusión a que las redes de acceso dan conectividad a Internet a suscriptores. En este contexto, el término suscriptor hace referencia a los hogares de usuarios de Internet que disponen una línea de acceso a la red con el proveedor del servicio. No obstante, es importante resaltar que dentro de un hogar pueden cohabitar un conjunto de usuarios que utilizan el mismo acceso a Internet.

La figura 4.7 muestra el diagrama del modelo de red de acceso utilizado en la metodología, teniendo en cuenta que dentro de un hogar existe un número de usuarios que también disponen de acceso a Internet. El objetivo de la metodología propuesta en este capítulo es la de obtener el rendimiento asociado al ancho de banda requerido por los usuarios a lo largo del tiempo $BW_{REQ}(t)$ en función de la demanda de tráfico de los M usuarios $(U_1, U_2, \ldots, U_M)$ correspondientes a los N suscriptores $(H1, H2, \ldots, H_N)$.

Al considerar la existencia de un conjunto de usuarios por suscriptor de red de acceso (hogar), se define un nuevo problema de acceso a recurso compartido:

- Cada usuario dispone de un enlace de comunicaciones hasta la puerta de enlace a Internet en el hogar, por ejemplo un router, que estará caracterizado por la

capacidad disponible $(D_1, D_2, \ldots, D_M)$, que depende de la tecnología de conexión, como por ejemplo, a través de una red WiFi o Ethernet.

- Los usuarios de la red doméstica pueden acceder a Internet de forma simultánea, por lo que competirán por el uso de un recurso compartido de red, es decir, la capacidad disponible desde el hogar hasta la red de acceso.

Este nuevo problema de acceso a recurso de red compartido podría ser analizado de forma análoga mediante la metodología expuesta a lo largo del capítulo. No obstante, se simplifica si se cumplen las siguientes hipótesis:

- La velocidad de conexión de la red doméstica no limita la velocidad que experimentan los usuarios al acceder a Internet (la velocidad de la tecnología de la red doméstica tiene unas prestaciones superiores al ancho de banda disponible en la red de acceso o la velocidad contratada con el operador).
- Los usuarios de un mismo hogar (o suscriptor de red de acceso) comparten el ancho de banda entre ellos siguiendo un esquema o algoritmo conocido, como por ejemplo, *max-min fairness*.

4.3.1.3. Parámetros del modelo

Como se ha descrito anteriormente, este modelo se define a partir de un conjunto de parámetros que detallan el escenario que se está analizando. Una vez que se seleccione el nivel de agregación o enlace de la red de acceso bajo estudio, se han de definir u obtener los siguientes parámetros del modelo de red:

1. **Número de suscriptores (N)**: número de hogares que tienen acceso a Internet y cuyo tráfico transita a través del enlace bajo estudio.
2. **Capacidades de suscriptores ($\{C_1, C2, \ldots, C_N\}$)**: las capacidades contratadas por los suscriptores con el proveedor del servicio u operador de red y que define las expectativas de velocidad que esperan obtener.
3. **Número de usuarios de Internet (M)**: cada suscriptor de red suele representar un hogar en el que pueden habitar un número de usuarios de Internet, que pueden hacer uso simultáneo de la red.
4. **Capacidades de usuarios de Internet ($\{D_1, D_2, \ldots, D_M\}$)**: de forma análoga a la anterior, cada usuario también se encuentra caracterizado por las prestaciones ofrecidas por la tecnología de conexión que utilice para acceder a la red.

El número y capacidades contratadas por los suscriptores son parámetros proporcionados por la red de acceso que quiere ser analizada. Estos datos suelen ser proporcionados

por el proveedor del servicio de Internet u operadores de Internet. En caso de no disponer de las capacidades contratadas por los suscriptores, esta información puede ser derivada de alguna fuente estadística, como por ejemplo en el caso de España, el INE o la Comisión Nacional de los Mercados y la Competencia (CNMC).

El número de usuarios de Internet por suscriptor a la red de acceso es un dato que puede obtenerse fácilmente a partir de estudios socio-demográfico que caracterice los hogares de una zona geográfica determinada e indique el número de miembros de la familia. A partir de esta información, gracias a la caracterización de usuarios de Internet, se pueden filtrar aquellos individuos que residan en el hogar pero que puedan no ser usuarios de Internet.

En referencia a las capacidades disponibles para cada usuario de Internet, se pueden realizar diferentes enfoques para confirmar si las tecnologías de acceso a la red doméstica suponen una limitación en las prestaciones de red que perciben los usuarios. Una simplificación consiste en asumir que no existe limitación de ancho de banda de red debido a la tecnología de acceso utilizada por los usuarios, por ejemplo a través de una conexión Ethernet o Wi-Fi.

4.3.2. Modelo de tráfico de red

En la sección 4.2.4 de este capítulo se presenta el modelo de estimación de rendimiento de un enlace en función de las fuentes de tráfico que agrega. En este contexto, estas fuentes caracterizan todo el tráfico que se genera a partir de la superposición de los usuarios del hogar y de las aplicaciones que estén siendo utilizadas.

4.3.2.1. Superposición de tráfico de usuarios

Las fuentes de tráfico a nivel de suscriptor son consideradas como fuentes de tipo ON/OFF, cuyos tiempos de actividad e inactividad vienen determinados por la superposición de los tiempos de actividad e inactividad de los usuarios de Internet que residen en el hogar del suscriptor de la red de acceso.

En la figura 4.8 se muestra la superposición de la actividad de uso de red de N usuarios de Internet, la cual define a su vez los patrones de actividad e inactividad a nivel de un suscriptor i. Estos periodos de actividad (ON) e inactividad (OFF) a nivel de suscriptor es equivalente al modelo de fuente de tipo ON/OFF definido en la sección 4.2.4 y representado en la figura 4.5.

En la figura anterior se ilustra únicamente el estado de actividad de los usuarios, activo (ON) o inactivo (OFF). No obstante, la duración de estos estados de actividad dependerán de cómo se reparta el ancho de banda disponible (a nivel de suscriptor de red de acceso) entre los usuarios.

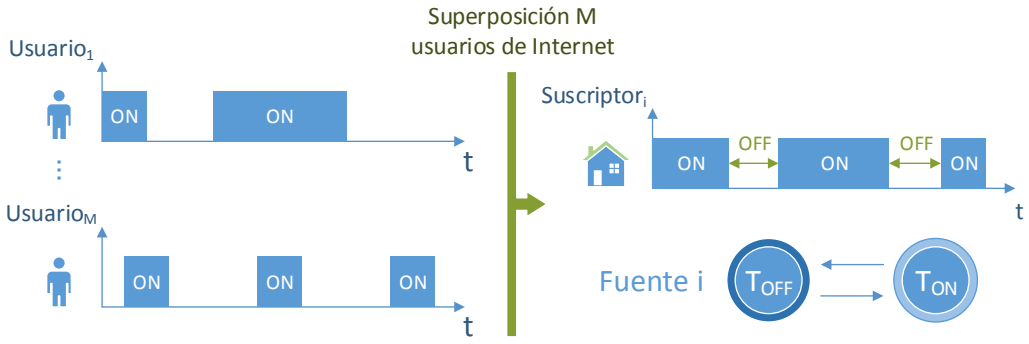

Figura 4.8: Diagrama de superposición de actividad de M usuarios de Internet

4.3.2.2. Mezcla de tráfico de aplicaciones

El componente elemental del modelo de tráfico presentado es la aplicación de Internet. En la sección 2.2.4 se introduce la necesidad de identificar aquellas aplicaciones de Internet representativas y con mayor impacto en la demanda global de Internet. Además de considerar sólo aquellas aplicaciones de mayor impacto en la demanda de tráfico, también se define un conjunto de aplicaciones de forma abstracta para que éstas puedan incluir una gran variedad de actividades realizadas en Internet (ecuación 4.5). Por ejemplo, el caso de las aplicaciones de redes sociales es un ejemplo de una actividad que podría incluirse en una aplicación de navegación web.

$$\text{Aplicaciones} = \{A_1, A_2, \ldots, A_L\} \tag{4.5}$$

El conjunto de aplicaciones consideradas para la mezcla de tráfico y, por tanto, para la estimación de demanda de tráfico de Internet, se clasifican mediante dos tipos de aplicaciones en función de la naturaleza de la actividad del usuario que la consume:

- *Foreground*: aplicaciones que se utilizan en primer plano por parte del usuario, es decir, los consumidores de este tipo de aplicaciones le prestan una atención prácticamente exclusiva a las mismas (por ejemplo, ver un video)

- *Backgound*: aplicaciones que se ejecutan en segundo plano y que permiten que los usuarios puedan estar utilizando otras aplicaciones a la vez (por ejemplo descargarse una película)

Esta clasificación modela la actividad de un usuario de Internet que puede tener un conjunto de aplicaciones de Internet generando tráfico en segundo plano, mientras

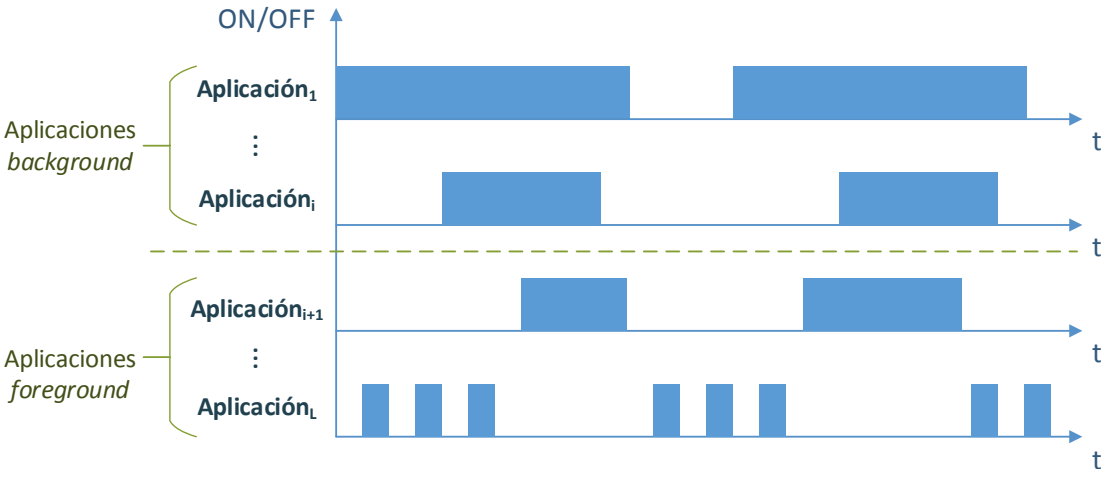

Figura 4.9: Diagrama de actividad de tipos de aplicaciones de usuario de Internet

utiliza otra aplicación que requiere toda su atención (en primer plano). A pesar de que pueden existir escenarios reales, donde un usuario se encuentre consumiendo dos aplicaciones de tipo *foreground* de forma simultánea (por ejemplo ver una película en streaming mientras realiza navegación web), este fenómeno no es considerado en el modelo presentado. En la ecuación 4.6 se divide el conjunto de aplicaciones en dos subconjuntos correspondientes a las aplicaciones *background* $(A_1, A_2, \ldots, A_i)$ y a las aplicaciones de tipo *foreground* $(A_{i+1}, A_{i+2}, \ldots, A_L)$.

$$\begin{aligned} \text{Aplicaciones} &= Aplicaciones_{BG} \cup Aplicaciones_{FG} \\ &= \{A_1, A_2, \ldots, A_i\} \cup \{A_{i+1}, A_{i+2}, \ldots, A_L\} \end{aligned} \tag{4.6}$$

La figura 4.9 muestra la actividad de un usuario de Internet que utiliza ambos tipos de aplicaciones a lo largo del tiempo. Se puede apreciar como las aplicaciones de tipo *foreground* no se solapan en el tiempo, mientras que las aplicaciones *background* se encuentran haciendo uso de la red independientemente del resto de aplicaciones. Es importante resaltar que a mayor número de aplicaciones activas simultáneas, los periodos de actividad se verán influenciados al disponer de un ancho de banda menor para cada aplicación.

4.3.2.3. Superposición de tráfico de aplicaciones

Análogamente a la superposición de tráfico de usuarios, los periodos de actividad (ON) e inactividad (OFF) de un usuario se ven definidos por la superposición de los tiempos de actividad de las aplicaciones que estén siendo utilizadas.

El tráfico de las aplicaciones de Internet consideradas, se modela mediante modelos

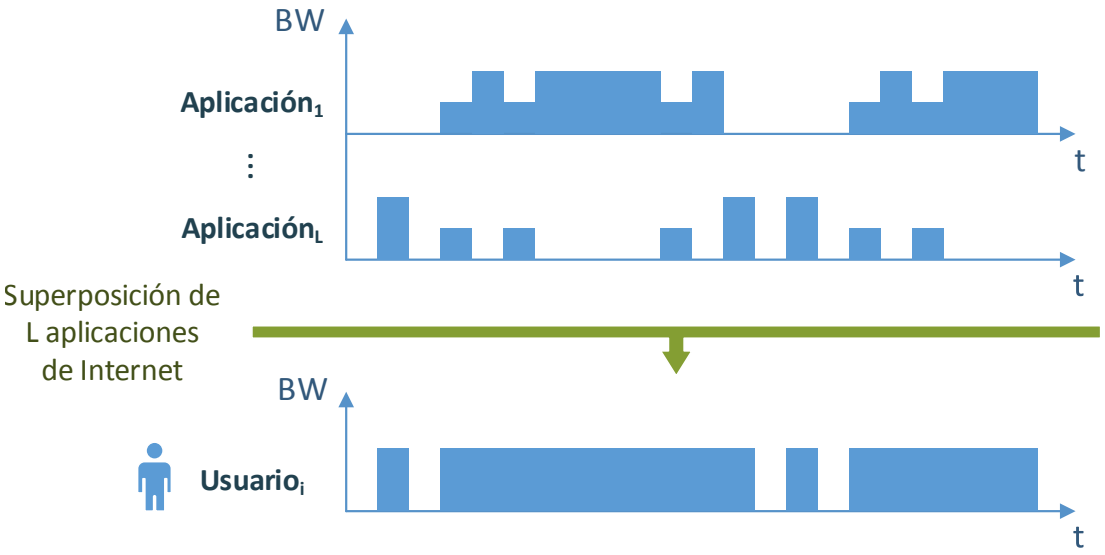

Figura 4.10: Diagrama de superposición de actividad de L aplicaciones de usuario

de fuente de tráfico de tipo ON/OFF. Así pues, para poder superponer los periodos de actividad (ON) e inactividad (OFF) de las aplicaciones consumidas por los usuarios, es necesario conocer o definir un modelo de tráfico específico para esa aplicación. Estos modelos de tráfico dependen íntimamente del tipo de aplicación de Internet que representen, pudiendo estar definidos con multitud de parámetros y dependiendo íntimamente de cómo se encuentren implementadas e incluso de los protocolos de comunicaciones subyacentes (por ejemplo, TCP o UDP).

En la figura 4.10 se presenta un ejemplo de superposición de actividades de diferentes aplicaciones de un usuario, cuya superposición define los periodos de actividad (ON) e inactividad (OFF) a nivel de usuario. De forma similar a la superposición de tráfico de usuarios, cuando L aplicaciones generan o consumen tráfico de forma simultánea, el ancho de banda disponible de usuario ha de dividirse entre estas aplicaciones siguiendo algún tipo de esquema de compartición.

Debido a los diferentes niveles de superposición de tráfico de aplicaciones y usuarios, los tiempos de actividad a nivel de suscriptor se definen a partir de los tiempos de actividad de los usuarios de Internet, que a su vez dependen los modelos de tráfico de las aplicaciones. La consecuencia lógica de esta dependencia a varios niveles de abstracción, es que el rendimiento de una aplicación, en términos de velocidad y tiempos de actividad (ON), depende del resto de aplicaciones activas en ese mismo usuario y la actividad de otros usuarios que residan en su mismo hogar.

De forma análoga a la superposición de usuarios, los tiempos de actividad de las aplicaciones dependerán de cómo se divida el ancho de banda disponible para el usuario

entre las aplicaciones que se encuentren activas. En este contexto existen multitud de posibilidades de repartición de ancho de banda entre usuarios y aplicaciones, pudiendo existir diferentes enfoques muy diferentes entre sí, como por ejemplo, máxima equidad entre usuarios y aplicaciones, o imponer prioridades de asignación de ancho de banda en función del tipo de usuario o del tipo de aplicación (con requisitos de red más exigentes).

4.3.2.4. Parámetros del modelo

Como se ha descrito anteriormente, este modelo se descompone en diferentes niveles de abstracción que sirven para definir el modelo global a partir de los niveles inferiores hasta el modelo global de tráfico (correspondiente a un suscriptor de la red de acceso). Los parámetros necesarios para definir el modelo de tráfico necesario para la metodología de estimación de demanda presentada en este capítulo son los siguientes:

1. **Selección de conjunto de aplicaciones ($\{A_1, A_2, \ldots, A_L\}$)**: se eligen un conjunto de aplicaciones representativas respecto a la demanda global de Internet. Se modelan las diferentes aplicaciones en función de si son de tipo *foreground* o *background*.

2. **Modelos de tráfico ON/OFF de aplicaciones de Internet**: se seleccionan o desarrollan modelos de tráfico basados en fuentes de tipo ON/OFF que caractericen el tipo de demanda de tráfico de las aplicaciones de Internet consideradas.

3. **Esquema de asignación de recurso de red entre usuarios y/o aplicaciones**: se selecciona un esquema de reparto de recursos de red que defina cómo se divide el ancho de banda asignado a cada usuario y/o aplicación.

Existen multitud de estudios y análisis a nivel global que identifican el conjunto de aplicaciones que más contribuye a la demanda de tráfico de Internet. A partir de estas referencias bibliográficas se puede escoger un conjunto adecuado y representativo de aplicaciones de Internet que sean válidas para la estimación de demanda de tráfico de Internet en una red de acceso. En la sección 2.2.4 se hace una revisión de algunos de estudios y se propone un conjunto de aplicaciones representativas: navegación web, compartición de ficheros, video sobre Internet y juegos en red.

Para cada aplicación del conjunto de aplicaciones representativas del tráfico de Internet, se ha de definir o escoger de la literatura, un modelo de fuente de tráfico. A pesar de que pueden existir muchos tipos de modelos basados en enfoques matemáticos diversos, esta tesis doctoral utiliza modelos basados en fuentes de tipo ON/OFF.

El último parámetro a especificar es el esquema de asignación de recursos de red entre usuarios y/o aplicaciones del modelo. Como se ha mencionado, existen diversos esquemas que tienen en cuenta diferentes tipos de usuarios (prioridad de algunos tipos de usuarios)

o incluso las prestaciones necesarias para cada tipo de aplicación (servicios diferenciados). En escenarios reales los protocolos de transporte utilizados (TCP, UDP, ...), así como, las propias implementaciones de las aplicaciones, juegan un papel importante la asignación de ancho de banda [Fred et al., 2001]. Con intención de simplificar el análisis de cómo se asigna el ancho de banda entre flujos de aplicaciones, se puede optar por escoger un esquema de compartición de ancho de banda sencillo, como por ejemplo, el *max-min fairness* [Bertsekas et al., 1992]. Este esquema tiene como objetivo maximizar las los anchos de banda asignados en los enlaces que llegan a los suscriptores y usuarios, es decir, el ancho de banda se reparte de forma justa primero entre usuarios y posteriormente entre aplicaciones.

4.3.3. Modelo de perfiles de usuario y aplicaciones

El modelo de perfiles de usuario y aplicaciones constituye el último componente del método de estimación demanda de tráfico descrito a lo largo de este capítulo. Mediante este modelo se define cómo los usuarios hacen uso de un conjunto de aplicaciones (definidas en la sección 4.3.2) sobre una red de acceso (descrita mediante el modelo de la sección 4.3.1).

En primer lugar, el modelo define una población de estudio representativa de los usuarios de la red de acceso, a partir de la cual se identifican y extraen un conjunto de perfiles de usuario. Para cada uno de estos perfiles, se define un conjunto de indicadores que caracterizan el uso de las aplicaciones y la actividad de usuario en la red para cada perfil de usuario.

4.3.3.1. Perfiles de usuario de la red

Este modelo ante la necesidad de caracterizar a los usuarios de una red de acceso desde un punto de vista de patrones de consumo de aplicaciones. Así pues, en primer lugar se ha de definir una población de estudio que se corresponda con los individuos que hagan uso de la red de acceso que se está analizando mediante la metodología de estimación de demanda de tráfico.

Población de estudio. La población de estudio tiene que representar en características al conjunto de usuarios de la red de acceso en términos de hábitos y comportamientos de consumo de Internet y sus aplicaciones. Para obtener esta población de estudio válida para la estimación de demanda de tráfico, se hacen corresponder las variables socio-demográficas del conjunto de usuarios de la red de acceso con los del conjunto de individuos que forman la población de estudio.

Esta correspondencia entre variables socio-demográficas y características de consumo de aplicaciones de Internet se fundamenta en los estudios científicos revisados en la

sección 2.1 y en el modelo conceptual desarrollado en el capítulo 3 de esta tesis doctoral. Es importante que la población de estudio caracterice correctamente aquellas variables socio-demográficas (edad, densidad de población de la zona geográfica, nivel de estudios, ...) que influyen en la adopción de las TICs y, especialmente, en los hábitos de consumo de Internet.

Identificación de perfiles de usuario. Una vez definida la población de estudio, se procede a identificar a un conjunto de perfiles de usuario de Internet que existen a partir de los diferentes patrones de uso y consumo de servicios. De esta forma se pueden clasificar a los usuarios de la red de acceso en K perfiles de usuario de Internet, conociendo además la probabilidad de pertenencia a cada perfil de usuario (p_i).

$$\begin{aligned} &\text{Perfiles de usuario} = \{P_1, P_2, \ldots, P_K\} \\ &\text{Prob. pertenencia a perfil} = \{p_1, p_2, \ldots, p_K\} \end{aligned} \tag{4.7}$$

Para la extracción de este conocimiento se pueden emplear diversas técnicas, como por ejemplo, el análisis de conglomerados a partir de datos socio-demográficos. Este enfoque, detallado a lo largo del capítulo 3 Caracterización de usuarios de Internet, extrae un conjunto de perfiles de usuarios de Internet a partir de variables que indican hábitos de consumo de actividades realizadas en Internet.

4.3.3.2. Caracterización de actividad de usuario

El primer componente del modelo es la caracterización de la actividad de usuario para cada perfil identificado en un periodo de tiempo determinado. Este modelo define la actividad de un perfil de usuario mediante dos características relativas a las conexiones que realizan a lo largo de un periodo de tiempo determinado:

1. Probabilidad de conexión: $pc_1, pc_2, \ldots, pc_K$
2. Tiempo de conexión medio: $tc_1, tc_2, \ldots, tc_K$

A partir de estos dos datos se puede estimar cuántos usuarios se encuentran conectados en un momento determinado y la duración media de las conexiones. A pesar de que el periodo puede variar en función de los datos estadísticos de entrada, el periodo más habitual en la literatura es un día.

Periodo de máxima demanda de tráfico. Si el objetivo de la aplicación de la metodología es dimensionar o establecer reglas de dimensionado de la red de acceso, el periodo de máxima demanda de tráfico ha de tenerse en cuenta. Durante este periodo los usuarios hacen un uso más intensivo de la red, por lo que la demanda de tráfico

es máxima. Desde un punto de vista conceptual, este periodo de tiempo es similar al término de *hora cargada*, definido en redes de telefonía tradicional.

Los indicadores de actividad de usuario pueden encontrarse expresados a lo largo de diferentes periodos de tiempo, como por ejemplo, la probabilidad de conexión de conexión de un usuario a lo largo de un día. En este caso, se propone el uso de la proporción de concurrencia de usuarios, ya que define el porcentaje máximo esperado de usuarios simultáneos durante el periodo de máxima demanda de tráfico en la red de acceso. El objetivo de esta definición es modelar unas condiciones de red adversas, de forma que se asegure la calidad del dimensionado de red.

4.3.3.3. Caracterización de uso de aplicaciones

El objetivo principal de la extracción de perfiles de usuario es la caracterización concisa del consumo de aplicaciones realizado por los perfiles de usuario de Internet. Para realizar esta labor se definen unos indicadores de uso de aplicaciones para cada perfil de usuario.

La caracterización de usuario suele realizarse a partir de información de naturaleza estadística, por lo que las variables que contienen información sobre los hábitos de consumo de Internet $(V_1, V_2, \ldots, V_S)$, no tienen por qué corresponderse directamente al conjunto de aplicaciones considerado anteriormente $(A_1, A_2, \ldots, A_L)$.

Las variables mencionadas hacen referencia a actividades realizadas en la red que requieren el uso o consumo de algunas de las aplicaciones consideradas en el modelo $(A_1, A_2, \ldots, A_L)$. La ecuación 4.8 muestra el conjunto de variables que representan a las actividades realizadas en la red por los usuarios.

$$\text{Variables sobre actividades en la red} = \{V_1, V_2, \ldots, V_S\} \tag{4.8}$$

Se parte de la suposición de que la información sobre el uso o consumo de una aplicación de Internet se encuentra contenida en las respuestas correspondientes al conjunto de variables $(V_1, V_2, \ldots, V_S)$. Por este motivo, se define un indicador de uso (U_i) de una aplicación i como una combinación lineal de un subconjunto de variables de actividades realizadas en la red ponderadas por unos factores de relevancia específicos para cada aplicación (4.9).

$$U(\text{Aplicación} = A_i) = U_i = \alpha_{1,i} \cdot V_1 + \alpha_{2,i} \cdot V_2 + \ldots + \alpha_{S,i} \cdot V_S, \forall i \in [1, L] \tag{4.9}$$

Estos factores de relevancia (α) dependen íntimamente de la cuestión sobre la actividad realizada en Internet y la aplicación i sobre la cual se quiere estimar su uso. Por esta razón, aquellas preguntas que no dispongan de información sobre el uso de la aplicación en cuestión tendrán un factor α nulo.

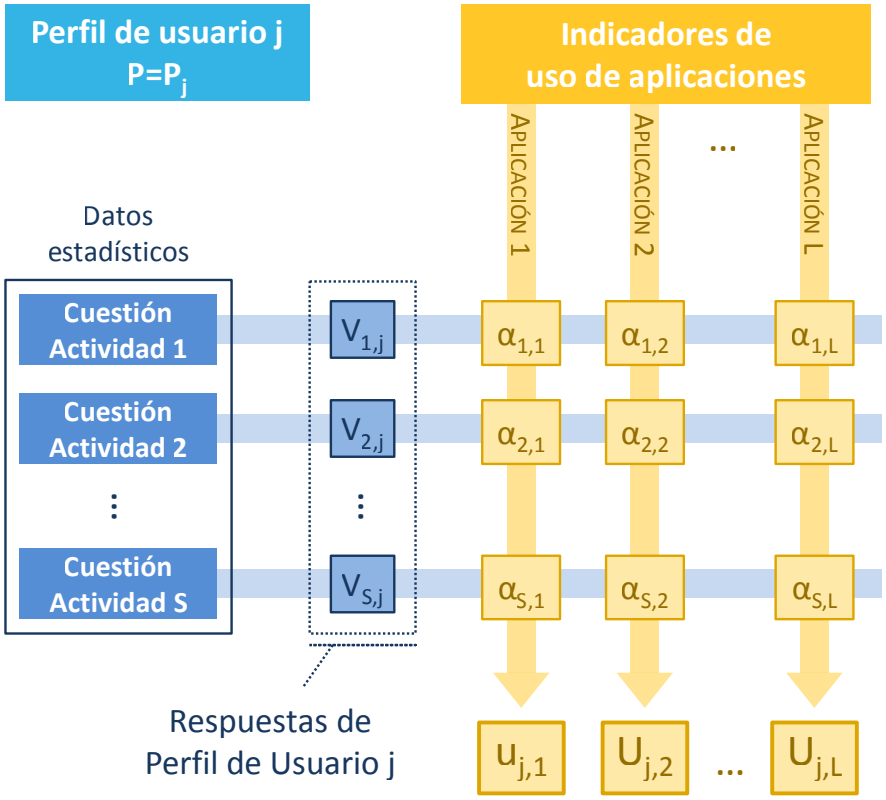

Figura 4.11: Diagrama del proceso de estimación de indicadores de uso de aplicaciones

En la ecuación 4.10 se particulariza la anterior ecuación para un perfil de usuario j. Los valores $\{v_{1,j}, v_{2,j}, ..., v_{S,j}\}$ se corresponden con las medias aritméticas de cada variable $\{V_1, V_2, \ldots, V_S\}$ de los individuos que conforman un perfil de usuario de Internet (P_j). La ecuación 4.10 define el uso de un perfil de usuario j para una aplicación i.

$$U_i(P = P_j) = U_{j,i} = \alpha_{1,i} \cdot V_{1,j} + \alpha_{2,i} \cdot V_{2,j} + \ldots + \alpha_{S,i} \cdot V_{S,j}, \forall i \in [1, L] \wedge \forall j \in [1, K] \tag{4.10}$$

En la figura 4.11 se muestran los componentes para la estimación de los indicadores de uso de aplicaciones para un perfil de usuario específico (P_j). Se aprecia que los factores de relevancia son independientes de los perfiles de usuario, y sólo dependen de la cuestión sobre la actividad y la aplicación de Internet.

Los factores de relevancia (α) han de ser ajustados debidamente para que los indicadores de uso de las aplicaciones representativas de la demanda de tráfico de Internet, sean coherentes y representen el consumo de tráfico de los usuarios de la red de acceso. Este proceso de ajuste depende las fuentes de información utilizadas para la metodología. En el capítulo 5 de esta tesis doctoral se presenta un ajuste de estos parámetros para aplicar la metodología a dos casos de estudio.

4.3.3.4. Parámetros del modelo

A continuación se enumeran los parámetros necesarios para poder definir el modelo de perfiles de usuario y aplicaciones, siguiendo las pautas anteriormente descritas:

1. **Perfiles de usuario ($P_1, P_2, \ldots, P_K$):** se extrae un conjunto de tipos de usuarios que difieren en los patrones y hábitos de consumo de aplicaciones de Internet.

2. **Probabilidad de pertenencia a perfil de usuario ($\{p_1, p_2, \ldots, p_K\}$):** se extraen las probabilidades de pertenencia a cada perfil de usuario de Internet.

3. **Indicadores de actividad de usuario (pc_i, tc_i):** se extrae un conjunto de variables que caracteriza la frecuencia o probabilidad de las conexiones de los perfiles de usuarios de Internet, así como la duración media de las mismas.

4. **Indicadores de uso de aplicaciones ($u_{i,j}$):** se extrae una matriz que caracteriza los patrones de consumo de las diferentes aplicaciones de Internet (j) para cada uno de los perfiles de usuario (i).

5. **Concurrencia de usuarios:** se define una proporción máxima esperada de usuarios simultáneos durante el periodo de máxima demanda de tráfico en la red de acceso.

En la figura 4.12 se muestra un diagrama del modelo donde se ilustra cada uno de los parámetros que lo componen. Como se aprecia en la figura, a partir de una población de estudio, se identifica un conjunto de perfiles de usuario de Internet ($\{P_1, P_2, \ldots, P_K\}$). Posteriormente, para cada perfil de usuario i, se extraen los indicadores de uso de aplicaciones ($u_{i,j}$) y de actividad de usuario (pc_i, tc_i).

4.3.4. Resumen: aplicación de metodología

A continuación se resume cómo aplicar la metodología de estimación de demanda, presentada en este capítulo, a un escenario concreto de una red de acceso residencial.

En primer lugar, se ha definir el escenario de red que se va a analizar a partir del modelo de red de acceso descrito en la sección 4.3.1, es decir, identificar y caracterizar todos los componentes de la de red de acceso analizada conforme a la figura 4.7, mediante los siguientes pasos:

1. Identificar el enlace que corresponde al nivel de agregación a considerar por el modelo, y cuyo ancho de banda requerido a lo largo del tiempo va a ser analizado (BW_{REQ}).

2. Identificar o definir el número de suscriptores (N) que va disponer de acceso a Internet a través del enlace, disponiendo de un conjunto de hogares ($H_1, H_2, \ldots, H_N$).

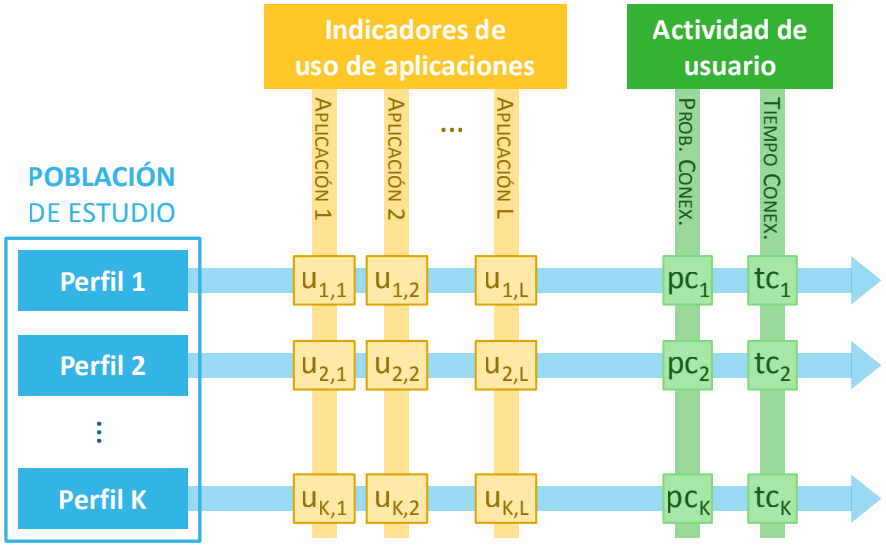

Figura 4.12: Diagrama del modelo de demanda de uso de aplicaciones

3. Caracterizar a las capacidades de los suscriptores a partir de información referente a la velocidad que tengan contrata $(C_1, C_2, \ldots, C_N)$.

4. Definir el número de usuarios (M), a partir de la estimación del número de usuarios por hogar, de forma que se disponga de un conjunto de usuarios de Internet $(U_1, U_2, \ldots, U_M)$.

5. Caracterizar a los usuarios mediante las capacidades disponibles en la conexión a Internet $(D_1, D_2, \ldots, D_M)$. Si se parte de la hipótesis que las tecnologías de conexión a la red doméstica no limitan el rendimiento del acceso, este paso es opcional.

Posteriormente, mediante el modelo de tráfico de red, se definen los modelos matemáticos que van a ser utilizados para estimar la demanda de tráfico de cada aplicación. Además, también se han de definir los enfoques o algoritmos utilizados para asignar los recursos de red entre usuarios y/o aplicaciones. Los pasos a seguir para utilizar este modelo a un caso de estudio son:

1. Definir un conjunto de aplicaciones de Internet que sea representativo de la demanda global de Internet $(A_1, A_2, \ldots, A_L)$.

2. Clasificar el conjunto de aplicaciones seleccionado en los diferentes tipos de aplicaciones considerados (*background* o *foreground*).

3. Seleccionar de la literatura o desarrollar un conjunto de modelos de tráfico basado en fuentes ON/OFF para cada aplicación de Internet.

4. Definir un esquema de asignación de recursos. Un enfoque muy sencillo consiste en suponer un esquema *max-min fairness*, de forma que se reparte equitativamente el ancho de banda, primero entre usuarios y después entre aplicaciones.

Por último, mediante el modelo de perfiles de usuario y aplicaciones, se caracteriza a los usuarios de la red a partir de los perfiles de usuario y sus correspondientes características de consumo de aplicaciones de Internet. Esta caracterización se realiza a partir de los siguientes pasos:

1. Identificar los perfiles de usuario en la población de estudio $(P_1, P_2, \ldots, P_K)$.
2. Extraer las probabilidades de pertenencia a los perfiles de usuario $(\{p_1, p_2, \ldots, p_K\})$.
3. Definir los indicadores de actividad de usuario (pc_i, tc_i) que caracterizan las conexiones de los perfiles de usuario de Internet.
4. Definir los indicadores de uso de aplicaciones $(u_{i,j})$ que caracterizan el consumo de los perfiles de usuario de Internet.
5. Definir un porcentaje de concurrencia de usuarios simultáneos que caracterice las condiciones de la red durante el periodo de máxima demanda de tráfico.

Llegado a este punto de la aplicación de la metodología, estos modelos aportan la base para poder analizar la actividad de un conjunto de usuarios de Internet en un escenario de red de acceso mediante una herramienta de simulación de eventos. En este contexto, se considera evento a un cambio en la actividad de una aplicación de usuario, es decir pasar de estar activa a inactiva, o viceversa. Este cambio de estado en una única aplicación, hace que el ancho de banda a repartir entre usuarios y aplicaciones tenga que ser reasignado. En definitiva, cada evento genera un cambio en los valores de ancho de banda disponible para cada aplicación y usuario.

Destacar que en esta metodología no se han especificado qué procedimientos o enfoques pueden ser llevados a cabo para derivar o estimar las siguientes probabilidades:

1. Probabilidad de que un usuario se encuentre activo durante un periodo determinado a partir de los indicadores de actividad y del porcentaje de concurrencia considerado.
2. Probabilidad de que un usuario activo se encuentre utilizando un conjunto de aplicaciones.

En el capítulo 5 se describen los enfoques que se han llevado a cabo para derivar esta información a partir de los datos extraídos de fuentes estadísticas. Este es el caso de los indicadores de uso de aplicaciones y los indicadores de actividad de usuario de Internet.

4.4. Conclusiones

En este capítulo se presenta una metodología de estimación de demanda de tráfico que aborda el problema asociado a la compartición de ancho de banda entre los usuarios de una misma red de acceso.

En primer lugar, se define el problema de recurso compartido de red y se describe el modelo teórico en el que se basa la metodología. Además, se introduce el concepto de GoS, a partir del cual se puede analizar cuantitativamente el rendimiento de una red de acceso en función del ancho de banda agregado.

En el resto del capítulo se describen los 3 modelos que conforman la metodología. El modelo de red de acceso define formalmente el escenario de red a partir de los diferentes niveles de agregación existentes en el mismo. El modelo de tráfico de red describe las superposiciones de actividades de usuarios y aplicaciones que han sido consideradas en la metodología. Además, este modelo también define los modelos de tráfico matemáticos que se utilizan para estimar las demandas de tráfico de los elementos que conforman el escenario de red. El modelo de perfiles de usuario y aplicaciones caracteriza cómo utilizan los usuarios de la red las aplicaciones de Internet consideradas mediante un conjunto de indicadores y parámetros extraídos de fuentes de información externas.

En el capítulo 5 de esta tesis doctoral se presenta la aplicación de esta metodología de estimación de demanda y dimensionado de redes de acceso a dos casos de estudio.

Capítulo 5

Aplicación a casos de estudio

5.1. Introducción

En este capítulo se presenta la aplicación de la metodología de estimación de demanda de tráfico de Internet, descrita en el capítulo 4, mediante el uso de un modelo de simulación de eventos discretos. El objetivo principal consiste en describir y detallar el proceso que se ha seguido, para aplicar la metodología mencionada a dos casos de estudio específicos de redes de acceso residenciales.

En primer lugar, este capítulo define un modelo de simulación, especificando sus principales características y objetivos. Posteriormente, se describe el desarrollo de una herramienta de simulación para aplicar la metodología de estimación de demanda de tráfico. A continuación, se realiza una validación del modelo de simulación con objeto de comprobar la aplicabilidad y calidad del mismo. Por último, se describe la aplicación de la metodología a los siguientes casos de estudio:

1. Análisis de rendimiento de una red de acceso utilizando los datos de tipologías de usuarios correspondientes al año 2012.

2. Análisis de rendimiento de la misma red de acceso utilizando el pronóstico de la evolución de la tipología de usuarios de Internet.

La aplicación de la metodología para una misma red de acceso empleando parámetros de entrada distintos, hace evidente la utilidad de la metodología y la herramienta de simulación desarrollada para el análisis del rendimiento y dimensionado de las redes de acceso.

5.2. Modelo de simulación

En la sección 4.2 se introduce el problema de acceso compartido a recurso de red que trata de abordar esta tesis doctoral, donde existe un conjunto de suscriptores que

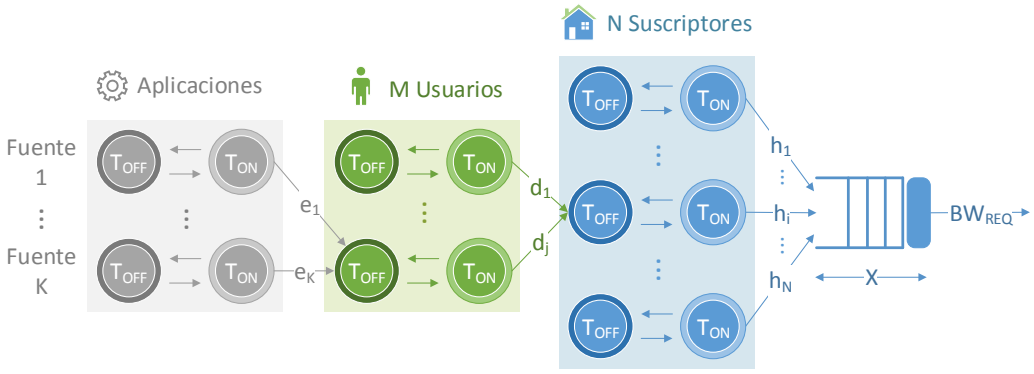

Figura 5.1: Modelo detallado de superposición de fuentes de tráfico de tipo ON/OFF

hacen uso de la red de acceso de forma simultánea. Los suscriptores demandan tráfico de la red a una tasa de bit igual a la velocidad de Internet contratada.

El problema reside en que la actividad de los mencionados suscriptores de red, no se modela mediante una única fuente de tráfico (figura 4.5), sino que viene determinada por las actividades de usuario que residen en un hogar determinado y, que a su vez, dependen de las aplicaciones que estén siendo utilizadas.

5.2.1. Descripción de fuentes de tráfico a modelar

En la figura 5.1 se muestra el modelo detallado de fuentes de tráfico de tipo ON/OFF resultante de la superposición de varios niveles de abstracción. El recurso de red compartido que se analiza es el ancho de banda requerido, de forma que los suscriptores de la red de acceso dispongan siempre de una tasa de bit ($\{h_1, \ldots, h_i, \ldots, h_N\}$) igual a la velocidad contratada con el proveedor de servicio de Internet. Las tasas de bit de los niveles de abstracción más bajos (por ejemplo en la figura, $\{d_1, \ldots, d_j\}$ y $\{e_1, \ldots, e_K\}$) vienen determinadas por la asignación equitativa del ancho de banda disponible (mediante un esquema de tipo *max-min fairness*). Los detalles del modelo de red de acceso y el modelo de tráfico se describen en la sección 4.3 de esta tesis doctoral.

En la figura 5.1 se ilustra cómo cada nivel de actividad viene definido por el nivel de actividad correspondiente al nivel de abstracción inferior (de derecha a izquierda):

1. Nivel de suscriptor de red acceso: este nivel se corresponde a los hogares con acceso a Internet, cuya actividad, modelada como una fuente ON/OFF, viene determinada por la superposición de actividades de los usuarios de Internet que utilizan el acceso a Internet del hogar. Existen N suscriptores de red acceso que pueden usar Internet de forma simultánea.

2. Nivel de usuario de Internet: son los individuos que residen en los hogares y que utilizan Internet. Su actividad de uso de Internet, también modelada como una fuente ON/OFF, viene determinada por los periodos de actividad de las aplicaciones que se encuentren utilizando. Existen M usuarios de Internet que pueden intentar acceder al recurso de red compartido.

3. Nivel de aplicación: este es el nivel más elemental del modelo, pues los periodos de actividad (ON) e inactividad (OFF) vienen determinados por un modelo de tipo ON/OFF para cada tipo de aplicación.

Es importante destacar que no todos los usuarios de Internet han de encontrarse activos y accediendo a la red de forma simultánea, al igual que no todos los usuarios han de utilizar la totalidad de las aplicaciones de tráfico consideradas en el modelo. Estos parámetros de los modelos se determinan de forma estocástica mediante distribuciones de probabilidad.

En resumen, el modelo consiste en múltiples fuentes heterogéneas de tipo ON/OFF (un tipo de fuente para cada tipo de aplicación de Internet considerada) que determinan mediante superposición de fuentes los dos niveles de fuentes ON/OFF superiores. Se asigna el ancho de banda de forma equitativa entre las entidades de cada nivel utilizando un esquema de *max-min fairness*. Este modelo resulta inabordable desde el punto de vista analítico por la heterogeneidad de las fuentes de tráfico y los niveles de superposición descritos. Por esta razón, se opta por analizar el rendimiento de la red de acceso mediante el uso de un modelo de simulación de eventos discretos.

5.2.2. Descripción del modelo de simulación

A continuación se detalla el modelo de simulación utilizado para la aplicación de la metodología de estimación de demanda de tráfico y dimensionado de redes de acceso. Se describen sus características y las principales decisiones de diseño para el desarrollo de la herramienta de simulación, utilizada en los casos de estudios presentados en este capítulo.

5.2.2.1. Características generales

El modelo de simulación presentado en este capítulo se basa en el análisis de eventos discretos a lo largo del tiempo. Este tipo de simulaciones se denominan DES. Cada evento representa una ocurrencia instantánea que genera un cambio en el sistema que está siendo modelado. Por esta razón, este modelo es de *estado discreto* ya que las variables analizadas representan valores discretos en el tiempo.

En el contexto de esta tesis doctoral, un evento representa un cambio en la actividad en algunas de las fuentes ON/OFF de las aplicaciones de algún usuario de Internet. Este

cambio, pasar de inactivo a activo o viceversa, produce una reasignación del recurso de red que está siendo compartido, en este caso el ancho de banda del enlace de la red de acceso analizada.

El modelo presentado en este capítulo es de tipo probabilístico (no-determinista), ya que cada simulación con los mismos parámetros de entrada producen resultados diferentes. Esto se debe a que el modelo se basa en funciones estocásticas y el uso de diferentes semillas para los generadores de números aleatorios utilizados en las simulaciones. Además, el modelo de simulación es estable y converge a una solución, siempre y cuando el tiempo de simulación sea lo suficientemente grande para permitir esta convergencia.

5.2.2.2. Objetivos de simulaciones

En esta sección se describen los objetivos que se persiguen mediante el modelo de simulación desarrollado y el proceso de diseño de escenarios de red, que posteriormente se aplican a los casos de estudio presentados en las secciones 5.4 y 5.5 respectivamente.

El objetivo principal del modelo de simulación es el análisis del rendimiento de un enlace de la red de acceso bajo unas condiciones de red y de tráfico específicas. A partir de la hipótesis de que los suscriptores de red siempre demandan un tráfico con una tasa de bit igual a la velocidad (o capacidad) contratada con el proveedor de servicio de Internet u operador, se analiza el ancho de banda requerido en el enlace a lo largo de un periodo de tiempo.

Cada simulación ha de recoger un conjunto de variables de forma que pueda calcular el rendimiento del enlace de la red de acceso a partir del GoS en función de los valores de ancho de banda requeridos lo largo del tiempo de simulación:

- Ancho de banda (BW_i) requerido en el enlace, para satisfacer la demanda de tráfico de la red derivada de los suscriptores y usuarios.
- Periodo de tiempo (T_{BW_i}) en el que se ha requerido un ancho de banda determinado (BW_i).

Además, las simulaciones que sigan este modelo también pueden recoger otras variables que contribuyan a la caracterización del escenario simulado y al análisis del rendimiento de la red de acceso.

5.2.2.3. Proceso de simulación

El proceso de simulación se puede definir mediante una secuencia de subprocesos. El diagrama mostrado en la figura 5.2 muestra el proceso en cadena que incluye 3 subprocesos:

Figura 5.2: Proceso en cadena correspondiente al modelo de simulación

- Creación de escenario de simulación: se define el escenario de simulación a partir de los modelos de red de acceso y modelo de perfiles de usuario y aplicaciones. Ambos modelos definen, por un lado la arquitectura de red de acceso en términos de entidades (suscriptores y usuarios) y sus capacidades, y por otro, los perfiles de usuarios y las aplicaciones de Internet que están siendo utilizadas en el escenario.

- Simulación de eventos discretos (DES): en esta fase se simulan las actividades de las entidades de la red (aplicaciones, usuarios y suscriptores de red de acceso), calculando para cada evento el ancho de banda requerido en el enlace de la red de acceso que están siendo analizado. También se extraen otras variables que permiten analizar y caracterizar el escenario y la simulación de red.

- Extracción de resultados: a partir de las variables extraídas en la fase anterior se realizan cálculos para conocer el rendimiento del enlace en función del ancho de banda requerido a lo largo del tiempo de simulación. También se extraen otros resultados y se presentan gráficamente.

Este proceso de simulación sienta las bases del desarrollo de una herramienta de simulación que pueda ser utilizada para la aplicación de la metodología de estimación de demanda de tráfico de usuario y dimensionado de red de acceso.

5.2.3. Selección de modelos de tráfico de aplicaciones

En esta sección se seleccionan los modelos de fuente de tráfico de tipo ON/OFF para un conjunto de aplicaciones que sean representativas de la demanda y consumo de tráfico de Internet, que ya han sido identificadas en la sección 2.2.4 de esta tesis doctoral: navegación web, compartición de ficheros, video sobre Internet y juegos en red. Estas aplicaciones son las más representativas y relevantes en relación a la demanda de tráfico que existe actualmente en Internet.

Es importante destacar que el conjunto de aplicaciones de Internet es un parámetro del modelo de simulación, que puede ser modificado o incluso puede evolucionar a medida que cambian los hábitos de consumo de tráfico en la red. Como se ha mencionado con anterioridad, para la aplicación de los casos de estudio sólo se considera el sentido descendente desde la red hacia los usuarios de Internet.

A continuación, se describe la selección de los modelos de fuente de tráfico para cada una de las aplicaciones mencionadas. Esta selección se fundamenta en el análisis del estado del arte de los modelos disponibles en la literatura, descrito en la sección 2.2. Se presta especial atención a los parámetros incluidos, la validez de representación de tráfico de red y la complejidad asociada a los modelos.

Navegación web. En la sección 2.2.5 se realiza un análisis de múltiples modelos que caracterizan la aplicación de navegación web. Se selecciona el modelo propuesto en [Pries et al., 2012] debido a que es un modelo que aborda algunas limitaciones que otros modelos no consideraban y que tiene en cuenta la popularidad de las páginas web existentes en la web.

En la ecuación 5.1 se define el tamaño total de tráfico que el usuario demanda, el cual viene determinado por la suma de objetos principales (S_M) junto con la suma de los tamaños de sus objetos en línea (S_{IN}). Las variables N_M y N_{IN} representan el número de objetos principales y el número de objetos en línea respectivamente. El tiempo de inactividad viene dado directamente por la variable T_OFF.

$$\begin{aligned} WEB_{ON} &= \sum_{i=1}^{N_M}\left(S_M(i) + \sum_{j=1}^{N_{IN}} S_{IN}(i,j)\right) \\ WEB_{OFF} &= T_{OFF} \end{aligned} \tag{5.1}$$

En la tabla 5.1 se muestran los parámetros de las distribuciones de probabilidad, los máximos y los mínimos establecidos en el modelo descrito en [Pries et al., 2012].

Compartición de ficheros. En la sección 2.2.6 se realiza un extenso análisis de todos los servicios disponibles en Internet que podrían englobarse dentro de esta categoría. Los servicios de mayor relevancia, en cuanto a la demanda de tráfico generada, se

Variable	Distribución	Parámetro 1	Parámetro 2	Max	Min
N_M	Lognormal	$\mu = 0,473844$	$\sigma = 0,688471$	212	1
S_M	Weibull	$\alpha = 28242,8$	$\beta = 0,814944$	8e6	-
N_{IN}	Exponencial	$\mu = 31,9291$	-	1920	1
S_{IN}	Lognormal	$\mu = 9,17979$	$\sigma = 1,24646$	8e6	-
T_{OFF}	Lognormal	$\mu = -0,495204$	$\sigma = 2,7731$	10000	-

Tabla 5.1: Parámetros del modelo de tráfico de aplicación de navegación web

corresponden con los servicios basados en P2P y en alojamiento de archivos en la red (*cyberlockers*). No obstante, después de esta revisión de la literatura se concluye con que no se ha definido ningún modelo de tráfico, que caracterice todos estos servicios en términos de tamaños de objetos y tiempos de inactividad.

Por esta razón y debido a que se conoce que la demanda de tráfico de usuario de este tipo de aplicaciones es muy intensa, en la aplicación de los casos de estudio de esta tesis doctoral, se opta por modelar el caso peor de demanda de tráfico de un usuario de este tipo de aplicación. El caso peor consiste en una simplificación absoluta del modelo, suponiendo que un usuario de esta aplicación siempre se encuentra demandando tráfico a una tasa de bit máxima y que por tanto nunca tiene periodos de inactividad (ecuación 5.2).

$$\begin{aligned} \text{FileSharing}_{ON} &\rightarrow \infty \\ \text{FileSharing}_{OFF} &= 0 \end{aligned} \tag{5.2}$$

Esta simplificación cobra especial relevancia cuando se modela el periodo de tiempo de máxima carga de tráfico en la red de acceso. Durante este periodo de tiempo es de suponer que un usuario de este tipo de aplicaciones pueda tener un gran número de ficheros en la cola de descarga de una aplicación. Por ello, el tráfico que demandaría a la red se encontraría cercano al caso peor que ha sido modelado. Este modelo también sería válido para los *cyberlockers* siempre y cuando el servidor no limite la velocidad de descarga y el contenido sea suficientemente grande.

Video sobre Internet. Los modelos de video sobre Internet analizados en la sección 2.2.7, se corresponden con aplicaciones basadas en diversos enfoques y características. Por esta razón, el tráfico generado por diferentes aplicaciones tiende a exhibir distintas características de tráfico. Además, muchos modelos de la literatura presentan importantes limitaciones debido a que sólo tienen en cuenta características temporales o porque han sido generados a partir de trazas de red sesgadas o bajo unas condiciones muy específicas, que no son aplicables a los casos de uso presentados en este capítulo.

Por estas razones, se opta por el uso del modelo descrito en [Zou et al., 2013] ya

que cuenta con una gran flexibilidad para escoger la calidad de video de streaming mediante la tasas de bit de codificación y que ha sido sido desarrollado para analizar el rendimiento en redes de comunicaciones.

Tomando como referencia que los usuarios de *YouTube* solicitan videos con tasas de bit medias que oscilan entre 632 y 908 Kbps [Zink et al., 2009], se selecciona un nivel de calidad del modelo de QL1, cuya tasa de bit de codificación es ampliamente superior a las anteriormente mencionadas. El ajuste de este parámetro se fundamenta en que para el análisis de la red de acceso y su dimensionado, es importante contemplar aquellas condiciones de red menos favorables, como puede ser el caso en que los usuarios demanden videos de alta calidad. De esta forma, también se tiene en consideración aquellos servicios de streaming de video que ofrecen calidades superiores a las de *YouTube*, como por ejemplo la plataforma *Netflix* donde se distribuyen contenidos de video de alta calidad.

El nivel de calidad utilizado ($QL1$) caracteriza un video con una tasa de bit de 1920 Kbps con una tasa de 25 frames por segundo (fps). El tamaño de frame S_{FRAME} viene determinado por una distribución de probabilidad de Pareto con los parámetros mostrados en la ecuación 5.3. A partir de estos parámetros se puede caracterizar el periodo de actividad a partir de la suma de tamaño de frames que demanda el usuario. El periodo de inactividad se modela como el tiempo restante que el usuario no requiere demandar más datos de video a la red.

$$\begin{aligned} \text{Video}_{ON} &= \sum_{i=1}^{fps} Pareto(\alpha = 1,2; k = 4800; m = 26100) \\ \text{Video}_{OFF} &= 1 - T_{\text{ON}} \end{aligned} \tag{5.3}$$

Si el tiempo actividad de video T_{ON} es el tiempo necesario para descargarse el video de 1 segundo de duración, el tiempo de inactividad (no se tiene en cuenta eventos como el buffering) se correspondería con el tiempo que el cliente de video no necesita descargarse más paquetes de video, es decir, la diferencia entre 1 segundo y el tiempo de actividad T_{ON}.

Juegos en red. En la sección 2.2.8 se realiza un análisis de los modelos de diferentes tipos de juegos en red más relevantes en la literatura. Se selecciona el modelo descrito en [Srinivasan et al., 2008], ya que caracteriza a los juegos de acción, los cuales tienen mayores requisitos de red, y por constituir unos de modelos más recientes en comparación con el resto de modelos analizados.

El modelo de demanda de tráfico seleccionado para este tipo de aplicaciones define los tamaños de paquetes de datos enviados por el servidor a usuarios a partir de una distribución de probabilidad de valor extremo con los parámetros mostrados en la

ecuación 5.4. El tiempo de inactividad viene determinado por otra distribución de valor extremo teniendo en cuenta el tiempo de actividad necesario para que el usuario se descargase los paquetes enviados por el servidor (T_{ON}).

$$\begin{aligned} \text{Gaming}_{ON} &= ExtremeValue(a = 330; b = 82) \\ \text{Gaming}_{OFF} &= ExtremeValue(a = 50; b = 4,5) \cdot 10^{-3} - T_{\text{ON}} \end{aligned} \tag{5.4}$$

5.2.4. Herramienta de simulación

Para utilizar el modelo de simulación anteriormente descrito, se ha optado por el desarrollo de un software secuencial programado en el lenguaje de programación propio de *Matlab* (lenguaje M). La decisión detrás de la elección de este Entorno de Desarrollo Integrado (IDE), reside la facilidad que brinda para manejar grandes cantidades de datos en forma matricial y las utilidades incluidas para la generación de gráficos y estadísticas. Estas razones hacen de *Matlab* un entorno ideal para desarrollar un prototipo de herramienta de DES que siga las pautas descritas anteriormente.

5.2.4.1. Consideraciones previas

La herramienta de simulación se ha desarrollado teniendo en cuenta un conjunto de consideraciones previas y con el objetivo de reducir la complejidad del análisis de la demanda de tráfico y dimensionado de redes de acceso.

La herramienta considera únicamente el sentido descendente de demanda de tráfico, es decir, en sentido desde la red de acceso hasta los usuarios. Se ha escogido realizar los ejemplos de aplicación de la metodología con el sentido descendente debido a que la mayoría de aplicaciones de Internet requieren más ancho de banda en sentido descendente. La única aplicación que requieren ancho de banda simétrico son las aplicaciones basadas en tecnologías P2P. No obstante, a pesar de la aparición de nuevos servicios que demandan anchos de bandas cada vez más simétricos, las aplicaciones más populares y más relevantes en cuanto al consumo de ancho de banda, siguen siendo altamente asimétricas [Pesovic and Sharpe, 2012].

En relación a cómo se reparte el ancho de banda disponible por los enlaces de la red entre suscriptores, usuarios de Internet y aplicaciones de usuarios, esta herramienta utiliza un esquema de asignación basado en el enfoque de máxima equidad entre entidades (*max-min fairness*). A pesar de que existen muchas estrategias de reparto de recursos de red, se ha escogido este enfoque debido a su simplicidad a la hora de desarrollar la herramienta de simulación. Otros enfoques, como por ejemplo, la priorización de algunos tipos de aplicaciones de Internet o de algunos tipos de usuarios, requerirían la inclusión de lógicas más complejas en la herramienta.

5.2.4.2. Parámetros de entrada

A continuación, se describen los diferentes parámetros de entrada necesarios para la ejecución de simulaciones mediante la herramienta. Estos parámetros pueden ser clasificados en dos grupos: parámetros de escenario de red de acceso y parámetros específicos de simulación.

Escenario de red de acceso. Los parámetros de escenario se corresponden con aquellos descritos en los modelos del capítulo 4. Algunos parámetros han sido obviados debido a que no son necesarios por las consideraciones anteriormente citadas.

Los parámetros de entrada correspondientes al escenario de la red de acceso analizada mediante la simulación son:

- Modelo de red de acceso
  - Número de suscriptores de la red de acceso
  - Distribución de probabilidad de capacidades de los suscriptores
  - Distribución de probabilidad que indica el número de usuarios por suscriptor
- Modelo de tráfico de red
  - Conjunto de aplicaciones representativas (background/foreground)
  - Modelos de fuente de tráfico ON/OFF de aplicaciones
- Modelo de perfiles de usuario y aplicaciones
  - Distribución de probabilidad de pertenencia a perfil de usuario (incluyendo un perfil de *No-usuarios* que representa a los individuos con acceso a Internet y que no utilizan la red)
  - Probabilidad y duración media de conexión de cada perfil de usuario (extraídos a partir de los indicadores de actividad de perfil de usuario)
  - Matriz de probabilidades de uso de aplicaciones representativas para cada perfil de usuario (extraídas a partir de los indicadores de uso de aplicaciones de perfiles)
  - Porcentaje de concurrencia de usuarios de Internet

Los modelos de demanda de tráfico de tipo ON/OFF correspondientes al conjunto de aplicaciones representativas han sido implementados mediante el lenguaje de programación de *Matlab* dentro de la propia herramienta de simulación. No obstante, han sido implementados como módulos independientes para que su modificación o reusabilidad sea fácilmente realizable.

Los parámetros de entrada anteriores, son definidos en un archivo de configuración que sigue una sintaxis específica de la herramienta para facilitar el uso, modificación y almacenamiento de diferentes escenarios de simulación.

Parámetros de simulación. Los parámetros de simulación son específicos de la ejecución de la herramienta, por lo que no tienen ningún tipo de relación con el escenario que está siendo simulado:

- Tiempo de simulación (en segundos)
- Tamaño de la muestra
- Tiempo de inicio de fuentes de tráfico (en segundos)
- Tiempo de inicio de análisis de resultados (en segundos)

El tiempo de simulación (T_{sim}) se corresponde con el periodo temporal que está siendo simulado, que no tiene por qué coincidir con el tiempo que tarda la herramienta en simular todos los eventos que transcurran durante este periodo.

Durante este tiempo de simulación, la herramienta recoge valores para aquellas variables que están siendo analizadas. El tamaño de muestra ($S_{muestra}$) se corresponde con el número de valores que van a ser recogidos por la herramienta. A partir de este parámetro se puede definir la frecuencia de muestreo de las variables analizadas en la simulación (ecuación 5.5).

$$f_{muestreo} = \frac{S_{muestra}}{T_{sim}} \tag{5.5}$$

El tiempo de inicio de fuentes de tráfico y el tiempo de inicio de análisis de resultados se utilizan para minimizar el periodo transitorio de la simulación, donde las fuentes de tráfico están siendo activadas y pueden generar resultados no válidos o de poca calidad. Por esta razón, el primer parámetro se utiliza para ir activando las fuentes de forma aleatoria a lo largo de un tiempo determinado, y el segundo para indicar el instante en el tiempo en el que se empiezan a recoger resultados de la simulación.

5.2.4.3. Resultados de la simulación

Los resultados de la simulación consisten en un conjunto de variables que indican el rendimiento de la red de acceso bajo un conjunto de condiciones impuestas mediante los parámetros de entrada descritos anteriormente.

Los resultados que se obtienen después de una ejecución de una simulación son los siguientes:

- GoS en función del ancho de banda requerido en el enlace (BW_{req})

- Distribución de usuarios activos simultáneos a lo largo del tiempo de simulación
- Distribución de aplicaciones simultáneas a lo largo del tiempo de simulación

Los datos obtenidos después de crear el escenario de simulación a partir de los parámetros de entrada, también pueden ser considerados resultados intermedios. Estos resultados, generados de forma aleatoria a partir de distribuciones de probabilidad, consisten en una caracterización de los usuarios de Internet a partir de las siguientes variables:

- Capacidad de enlace disponible de usuarios
- Perfil de usuario de Internet
- Tipos de aplicaciones de Internet

5.2.4.4. Descripción de componentes

A continuación, se describen los componentes desarrollados para la herramienta de simulación y que se encargan de proporcionar las funcionalidades necesarias para ejecutar las simulaciones siguiendo la descripción del modelo descrito en la sección 5.2.2. Además, la herramienta de simulación también implementa algunas funcionalidades adicionales:

- Guardar parámetros de entrada y resultados de simulaciones realizadas
- Cargar parámetros de entrada y resultados simulaciones realizadas
- Visualización de resultados mediante gráficas
- Guardar gráficas de resultados en diferentes formatos

En la figura 5.3 se muestra un diagrama simplificado de los componentes software desarrollados para la herramienta de simulación y que se encuentran escritos en el lenguaje M, propio de *Matlab*. Como se aprecia en la figura, existen 4 componentes principales con funcionalidades distintas y que componen la herramienta de simulación:

- Controlador y Graphical User Interface (GUI) (clase *simGUI*)
- Creación de escenario de simulación (clase *createScenario*)
- Ejecución de simulación (clase *bandwidthSim*)
- Visualización de resultados (clase *plotSimResults*)

Además, en la figura 5.3 también se aprecian dos paquetes que contienen otros componentes software:

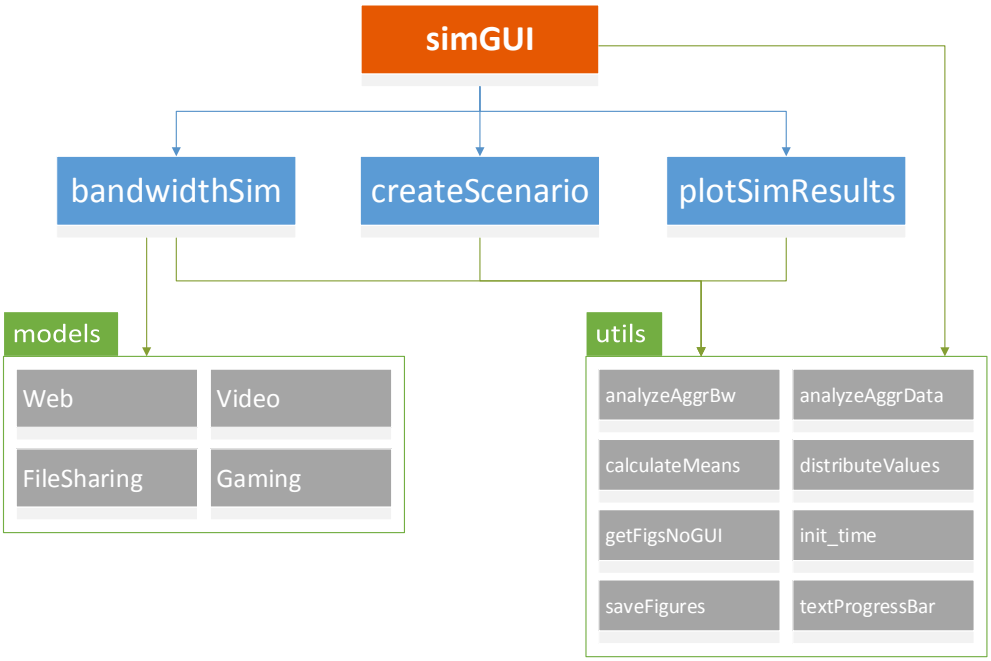

Figura 5.3: Diagrama de clases en lenguaje M de herramienta de simulación

- Modelos de tráfico de aplicaciones de Internet (paquete *models*)

- Clases que implementan funciones auxiliares (paquete *utils*)

Los modelos de tráfico de aplicaciones de Internet incluidos en la herramienta de simulación se corresponden con los seleccionados y descritos en la sección 5.2.3. Las implementaciones de los modelos de tráfico de aplicaciones siguen los modelos estocásticos definidos en la citada sección.

Las clases incluidas dentro del paquete *utils* son un conjunto de componentes software que implementan funcionalidades auxiliares y que son utilizadas en diferentes momentos de las ejecuciones de las simulaciones. Incluyen funcionalidades relacionadas con cálculos matemáticos (*analyzeAggrBw*, *analyzeAggrData*, *calculateMeans*, *distributeValues*, *init_time*) o con funciones asociadas con la GUI (*getFigsNoGUI*, *saveFigures*, *textProgressBar*).

A continuación, se introducen los 4 componentes principales de la herramienta con el objetivo de describir los pasos seguidos para realizar las simulaciones de los eventos que representan los cambios en la demanda de tráfico de las aplicaciones de los usuarios de Internet.

Controlador y GUI (*simGUI*) La clase *simGUI* constituye el componente software principal de la herramienta, pues controla el hilo de ejecución de las simulaciones y pone a disposición del usuario un conjunto de componentes gráficos para su interacción con la herramienta.

La GUI facilita al usuario la introducción de los parámetros de entrada anteriormente descritos, de forma que puede realizar ejecuciones de simulaciones de forma rápida y sencilla. Como se ha descrito anteriormente, la herramienta utiliza un archivo de configuración que incluye todos los parámetros propios de los modelos de red de acceso y de demanda de tráfico de usuarios.

En la figura 5.4, se aprecia como la GUI de la herramienta de simulación dispone de 3 áreas bien diferenciadas, donde el usuario puede realizar diferentes acciones o visualizar diferentes tipo de información:

- Parámetros de entrada: esta área cuenta con diferentes áreas de texto, donde puede introducir diferentes parámetros de entrada de escenario y de simulación.
- Operaciones: se incluyen diferentes botones que disparan diferentes funciones de la herramienta (ejecutar simulación, cargar y salvar simulación, representar resultados y guardar figuras).
- Resultados: área donde se presenta un esbozo de los resultados del escenario creado y de la simulación.

Como se ha mencionado anteriormente, este componente software también tiene funciones de controlador pues se encarga de lanzar y ejecutar funciones de la herramienta. Estas funciones se disparan a partir de los botones disponibles en el área de *Operaciones*.

Creación de escenario de simulación (*createScenario*) Este componente tiene como principal objetivo calcular o extraer las características del escenario de red que son necesarias para la ejecución de simulaciones. Todas estas características se derivan o calculan directamente de los parámetros de entrada especificados por el usuario en la herramienta.

A continuación se resumen los cálculos más relevantes realizados por este componente de la herramienta:

1. Calcular el número de usuarios por suscriptor de red (hogar)
2. Calcular la capacidad de cada suscriptor de red (hogar)
3. Asignar un perfil de usuario de Internet a cada usuario del escenario (se identifican aquellos individuos con acceso a Internet que no utilizan la red y forman parte del segmento de *No-Usuarios*)

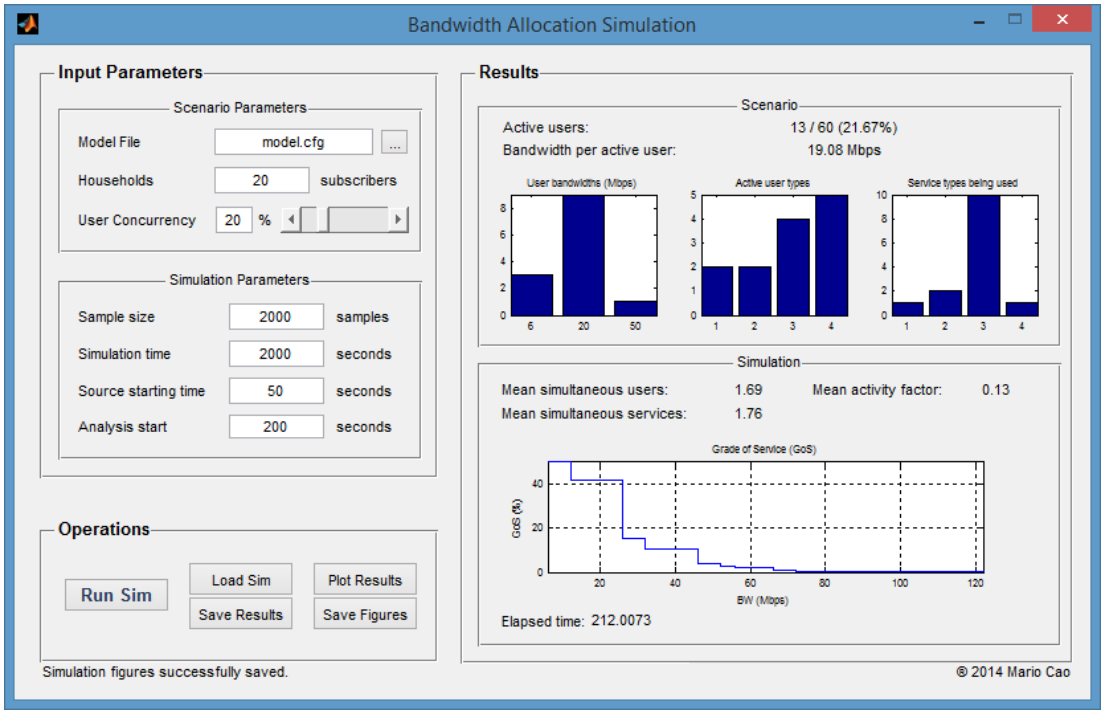

Figura 5.4: Interfaz gráfica de usuario de la herramienta de simulación

4. Calcular la actividad de usuario de Internet teniendo en cuenta los siguientes parámetros

   a) Concurrencia de usuarios que determina los usuarios conectados (activos)

   b) Aplicaciones utilizadas por usuarios activos

Todos los cálculos anteriores se realizan a partir de los parámetros de entrada de los usuarios y basándose en funciones de probabilidad especificadas por el usuario de la herramienta. Por este motivo, los escenarios definidos para cada ejecución de simulación pueden contener características ligeramente diferentes unos a otros.

Para determinar si un usuario de Internet se encuentra activo o inactivo se considera el porcentaje de concurrencia introducido por el usuario de la herramienta. Además, dado que la probabilidad y duración de conexión varía entre diferentes perfiles de usuario de Internet, se realiza un conjunto de cálculos que consideran estas variables para la estimación del número de usuarios activos de cada perfil.

En la ecuación 5.6 se define el porcentaje de concurrencia ($\%Conc$) como la proporción de usuarios activos, $N_{ACTIVOS}$ y los usuarios totales (incluyendo aquellos *No-Usuarios* con acceso a Internet), N_{TOTAL}. El numerador puede descomponerse como el número total de usuarios de un perfil (N_i) multiplicado por la probabilidad de que se

encuentren activos ($P_{ACTIVO,i}$).

$$\%Conc = \frac{N_{ACTIVOS}}{N_{TOTAL}} = \frac{\sum_{i=1}^{K} P_{ACTIVO,i} \cdot N_i}{N_{TOTAL}} \tag{5.6}$$

Se define la probabilidad que un perfil se encuentre activo en función de una constante β, de la probabilidad de conexión (pc_i) y del tiempo de conexión (tc_i) de un perfil (ecuación 5.7). Además, se expresa el número de usuarios de un perfil en función del número total de usuarios (N_{TOTAL}), de la probabilidad de que no sean *No-Usuarios* ($1 - p_0$) y de la probabilidad de pertenecer a un perfil (p_i).

$$N_{ACTIVOS} = \sum_{i=1}^{K} \beta \cdot pc_i \cdot tc_i \cdot N_i = \beta \cdot \sum_{i=1}^{K} pc_i \cdot tc_i \cdot p_i \cdot (1 - p_0) \cdot N_{TOTAL} \tag{5.7}$$

A partir de las ecuaciones 5.6 y 5.7, se obtiene el valor de β (ecuación 5.8), donde todas las variables son conocidas.

$$\beta = \frac{\%Conc}{1 - p_0} \cdot \frac{1}{\sum_{i=1}^{K} pc_i \cdot tc_i \cdot p_i} \tag{5.8}$$

En consecuencia, la ecuación 5.9 define la probabilidad de que un usuario de un perfil se encuentre activo en función de parámetros de entrada de la herramienta.

$$P_{ACTIVO,i} = \beta \cdot pc_i \cdot tc_i = \frac{\%Conc}{1 - p_0} \cdot \frac{pc_i \cdot tc_i}{\sum_{i=1}^{K} pc_i \cdot tc_i \cdot p_i} \tag{5.9}$$

A partir de la ecuación 5.9, la herramienta de simulación determina el estado de actividad de los usuarios de Internet en función del perfil de usuario al que pertenezcan, respetando el porcentaje de concurrencia especificado por el usuario de la herramienta.

Ejecución de simulación (*bandwidthSim*) El componente más importante de la herramienta es el que se encarga de realizar la simulación de las fuentes de tráfico de los usuarios que conforman el escenario de red.

En la figura 5.5 se muestra un diagrama de actividad que ilustra el proceso de simulación implementado en este componente software. La simulación de eventos consiste en un proceso iterativo que va actualizando variables de simulación y extrayendo resultados intermedios hasta que se satisface una condición que consiste en superar el tiempo de simulación introducido por el usuario de la herramienta.

La primera etapa del proceso de simulación consiste en en la ejecución secuencial de las 3 primeras actividades mostradas en el diagrama de actividad (figura 5.5):

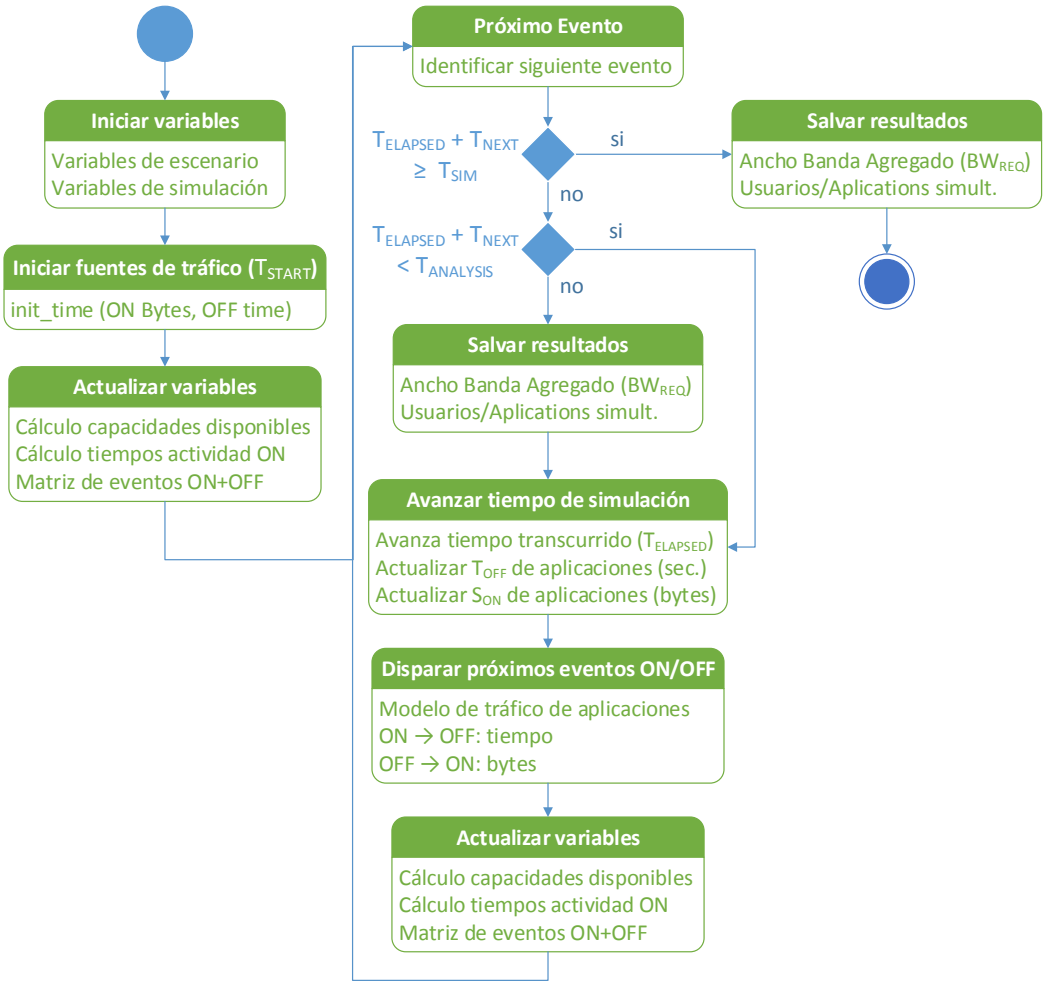

Figura 5.5: Diagrama de actividad de simulación de eventos

1. **Iniciar variables.** La primera actividad consiste en iniciar todas las variables necesarias para el entorno de simulación, en las que se incluyen las variables relacionadas con el escenario (suscriptores de red de acceso, usuarios, aplicaciones) y con las propias simulaciones (estado de actividad de usuarios y aplicaciones, bytes restantes de aplicaciones activas, tiempo restante de inactividad de aplicaciones).

2. **Iniciar fuentes de tráfico.** Las fuentes de tráfico pueden comenzar activas o inactivas. Durante esta actividad se caracterizan los periodos de actividad o inactividad durante un periodo inicial T_{START}, durante el cual se van a ir activando o desactivando siguiendo una función de distribución de probabilidad determinada (por ejemplo, una distribución uniforme).

3. **Actualizar Variables.** A continuación se procede a actualizar las variables con el objetivo de calcular la capacidad disponible para usuario de Internet (considerando que pueden haber varios usuarios en un mismo hogar) y para cada aplicación de Internet (considerando que pueden haber varias aplicaciones activas en un mismo usuario). Con estas capacidades para cada aplicación, se calculan los tiempos de actividad para las fuentes de tráfico activas (si esa capacidad se mantuviese indefinidamente), para poder generar así una matriz de eventos ON/OFF para todos los usuarios y aplicaciones. Esta matriz contiene los tiempos de cuándo se dispararían los eventos si las condiciones de red no cambiasen.

La segunda etapa se encuentra formada por un conjunto de actividades que se repiten hasta que se supera el tiempo de simulación total (T_{SIM}) especificado por el usuario de la herramienta en la ecuación 5.10a, donde $T_{ELAPSED}$ representa el tiempo transcurrido de simulación y T_{NEXT} es el tiempo necesario para el siguiente evento. Además, la herramienta no comienza a recolectar y analizar las variables de resultados mientras no se alcance un umbral $T_{ANALYSIS}$, definido por el usuario de la herramienta (ecuación 5.10b).

$$T_{ELAPSED} + T_{NEXT} \geq T_{SIM} \tag{5.10a}$$

$$T_{ELAPSED} + T_{NEXT} \leq T_{ANALYSIS} \tag{5.10b}$$

A continuación se describen las actividades de esta segunda etapa y cuya ejecución depende de las condiciones anteriormente descritas:

1. **Próximo Evento.** Se identifica el próximo evento y se calcula la marca de tiempo en la que acontecerá el próximo cambio de actividad de una o varias aplicaciones de Internet (T_{NEXT}).

2. **Salvar Resultados**. Se guardan los valores de las variables de ancho de banda

agregado requerido en el enlace de red de acceso y los valores de usuarios y aplicaciones simultáneos.

3. **Avanzar tiempo de simulación.** El tiempo de simulación transcurrido ($T_{ELAPSED}$) avanza el tiempo del siguiente evento identificado (T_{NEXT}).

   - Se reduce T_{NEXT} de los tiempos de inactividad de las aplicaciones inactivas en este momento de la simulación (T_{OFF}).
   - Se reduce los valores de *Bytes* restantes de las aplicaciones que se encontraban activas (S_{ON}) teniendo en cuenta las capacidades disponibles por aplicación y usuario.

4. **Disparar próximos eventos ON/OFF.** En esta actividad se dispara el evento o eventos correspondientes al tiempo del próximo evento T_{NEXT}. Se hace uso de los modelos de demanda de tráfico definido para cada una de las aplicaciones de Internet consideradas en la herramienta.

   - Si la fuente de tráfico se encontraba activa, ésta se desactiva durante un tiempo (T_{OFF}) determinado por el modelo de tráfico de la aplicación en cuestión.
   - Si la fuente de tráfico se encontraba inactiva, ésta se activa demandando un tráfico caracterizado por un valor de *Bytes* a partir del modelo de tráfico de la aplicación. El tiempo de actividad (T_{ON}) se define por el tiempo necesario para consumir esos datos determinado por la capacidad disponible para ese usuario y aplicación.

5. **Actualizar variables.** Análogo a la actividad descrita anteriormente con el mismo nombre, donde se guardan las capacidades, se calculan los tiempos de actividad y se genera la matriz con los tiempos de los próximos eventos.

Visualización de resultados (*plotSimResults*) Este componente se encarga de representar de forma gráfica los principales resultados de la ejecución de una simulación. Los resultados visualizados se clasifican en 3 categorías diferentes en función del tipo de variables que se muestran:

- Escenario de simulación: se muestran gráficas que caracterizan a los usuarios que se encuentran en el escenario a partir de la capacidad que tienen disponible, al perfil de usuario de Internet que pertenecen y las aplicaciones que están utilizando.

- Simultaneidad de usuarios y aplicaciones: se muestran gráficas con los porcentajes de usuarios y aplicaciones activos simultáneos a lo largo del tiempo de simulación.

- Rendimiento de la red de acceso: se muestra un gráfico del GoS en función del ancho de banda requerido por el enlace de la red de acceso que está siendo analizado.

5.3. Validación del modelo de simulación

Esta sección tiene como principal objetivo validar el modelo de simulación a partir de un experimento y asumiendo aquellas condiciones, que permitan comparar sus resultados con los de un modelo analítico.

En primer lugar, se describe el modelo analítico utilizado para la validación. Posteriormente, se detalla el escenario de simulación de red utilizado para que sus resultados puedan ser comparados con los extraídos del modelo analítico. Finalmente, se extraen las principales conclusiones de los resultados obtenidos y se analiza el alcance de la validación del modelo de simulación.

5.3.1. Modelo analítico

Se considera un escenario donde existen múltiples fuentes de tráfico ON/OFF que utilizan un recurso de red representado como una cola de un sistema de *burst scale queueing*, tal y como se describe en [Pitts and Schormans, 2001]. El modelo se basa en N fuentes de tráfico idénticas y que operan de forma independiente que utilizan una cola caracterizada por una capacidad de servicio C y un tamaño de búfer X.

En la ecuación 5.11 se define la tasa de bit media de una fuente de tráfico, m, a partir de las duraciones medias de actividad T_{ON} e inactivdad T_{OFF} de las fuentes, y de la tasa de bit, h, que usa una fuente cuando se encuentra activa .

$$m = h \cdot \frac{T_{ON}}{T_{ON} + T_{OFF}} \tag{5.11}$$

A partir de la tasas de bit anteriores, en la ecuación 5.12 se define el factor de actividad α, que además se corresponde con la probabilidad de que una fuente se encuentre activa.

$$\alpha = \frac{m}{h} = \frac{T_{ON}}{T_{ON} + T_{OFF}} \tag{5.12}$$

La condición para que exista necesidad de realizar *buffering* en el sistema se da cuando la suma de tasas de bit de fuentes activas es mayor que la tasa de servicio de la cola. Se define como N_0 el número de veces que cabe la tasa de bit máxima h en la capacidad de servicio C (ecuación 5.13).

$$N_0 = \frac{C}{h} \tag{5.13}$$

El número entero inmediatamente superior a $\lceil N_0 \rceil$, N_0, representa el mínimo número

de fuentes necesarias para que exista un evento de *buffering*. De forma inversa, el número entero inmediatamente inferior, $\lfloor N_0 \rfloor$, representa el máximo número de fuentes que pueden estar activas en el sistema sin que exista *buffering*.

A partir de estas definiciones, si se asume un modelo donde no existe posibilidad de eventos *buffering*, se puede calcular el factor de pérdidas de fuentes activas que no pueden ser atendidas.

Para una única fuente de tráfico, la probabilidad de que se necesitase *buffering* se podría calcular mediante la siguiente ecuación 5.14. Esta ecuación también puede verse como la división entre la tasa media de pérdida (*mean excess rate*) y la tasa de llegada media.

$$\text{Prob}\{\text{buffer}\} = \frac{\alpha(h-C)}{m} = \frac{\alpha(h-C)}{\alpha h} = \frac{h-C}{h} \tag{5.14}$$

Para el caso de múltiples fuentes de tráfico, se necesita calcular la probabilidad de que existan n fuentes activas mediante la distribución binomial (ecuación 5.15).

$$p_n = \frac{N!}{n!(N-n)!} \cdot \alpha^n \cdot (1-\alpha)^{N-n} \tag{5.15}$$

Teniendo en cuenta sólo aquellas fuentes que excedan la capacidad de servicio, es decir, $N_0 < n \leq N$, se calcula la tasa media de pérdida (*mean excess rate*) según la ecuación 5.16.

$$\text{Mean excess rate} = \sum_{n=\lceil N_0 \rceil}^{N} p_n \cdot (nh - C) \tag{5.16}$$

Conociendo que la tasa de llegada media es Nm, la probabilidad de que se necesite de un evento de *buffering* se calcula como la proporción entre la tasa media de pérdida entre la tasa media de llegada (ecuación 5.17). Además, utilizando las ecuaciones 5.12 y 5.13, la ecuación queda como:

$$\text{Prob}\{\text{buffer}\} = \frac{\sum_{n=\lceil N_0 \rceil}^{N} p_n \cdot (nh - C)}{Nm} = \frac{\sum_{n=\lceil N_0 \rceil}^{N} p_n \cdot (n - N_0)}{N\alpha} \tag{5.17}$$

5.3.2. Definición de escenario de simulación

Como se ha comentado en la sección 5.2, el modelo de simulación descrito no es abordable desde un punto de vista analítico. No obstante si se imponen un conjunto de suposiciones, los resultados obtenidos a partir de una ejecución de simulación podrían ser comparados con los del modelo analítico.

El modelo analítico, descrito anteriormente, parte de la suposición de que todas las fuentes de tráfico son idénticas, independientes entre sí y con una tasa de bit cuando están activas de valor h. Las principales implicaciones que tendría esta suposición en el modelo de simulación son las siguientes:

- **Fuentes de tráfico ON/OFF homogéneas.** Sólo se considera una única fuente de tráfico de aplicación de Internet para que éstas sean idénticas.
- **Capacidades homogéneas de suscriptores de red de acceso.** Debido a que la tasa de fuente activa es h en el modelo analítico, las capacidades que disponen los suscriptores a la red ha de ser igual para todos ellos.
- **Único usuario de Internet por suscriptor.** La inclusión de un nivel de agregación en el modelo de simulación tendría como consecuencia que las tasas de fuentes activas pudieran no ser siempre iguales a h (por ejemplo si hay dos usuarios activos simultáneos en un mismo hogar).

Al considerar una única aplicación y un único usuario de Internet por suscriptor, la definición de perfiles de usuarios de Internet carece de sentido, por lo que no se consideran diferentes tipos de usuarios en el escenario.

En el modelo analítico se define una tasa o capacidad de servicio C que caracteriza a la cola donde se agrega el tráfico de las fuentes. En el caso del modelo de simulación no se requiere definir la capacidad del enlace de red de acceso analizado, que se correspondería con la cola del modelo analítico. Para poder comparar resultados entre ambos modelos sólo se requiere fijar una capacidad de servicio (o de enlace) determinada para poder aplicar las ecuaciones del modelo analítico.

Por lo tanto, para poder comparar ambos modelos, se diseña un escenario de simulación que tenga en cuenta todas las implicaciones anteriores y que facilite el análisis de los resultados. Los parámetros de escenario utilizados en este contexto se definen a continuación:

- Número de suscriptores: $N = 50$ (1 usuarios por suscriptor)
- Concurrencia de usuarios: 100 %
- Capacidades de suscriptores: $h = 10Mbps$
- Modelo de tráfico de aplicación:
  - Bytes a consumir durante tiempo de actividad: $S_{ON} = exprnd(\frac{10^6}{8})$
  - Tiempo de inactividad: $T_{OFF} = exprnd(1)$

El modelo de tráfico utilizado para este escenario se ha escogido para que el factor de actividad α fuese cercano al 50 % para capturar una mayor simultaneidad de fuentes de tráfico y poder comparar este fenómeno con el modelo analítico con facilidad. Los periodos de actividad e inactividad vienen determinados por número aleatorios ($exprnd(\mu)$) generados a partir de una función exponencial (ecuación 5.18).

$$f(x|\mu) = \frac{1}{\mu} \cdot e^{\frac{-x}{\mu}} \tag{5.18}$$

Además, se utilizan parámetros de simulación suficientemente altos para que aseguren la convergencia de resultados de la ejecución y puedan ser comparados con los valores extraídos del modelo analítico:

- Tiempo de simulación: $T_{SIM} = 29200$ segundos (8 horas + 400 segundos)
- Tiempo de inicio de fuentes de tráfico: $T_{START} = 50$ segundos
- Tiempo de inicio de análisis: $T_{ANALYSIS} = 400$ segundos

5.3.3. Comparación de resultados

A continuación se presenta la comparación entre los resultados obtenidos del modelo de simulación, teniendo en cuenta las condiciones anteriores, y el modelo analítico. Para esta comparación se consideran las siguientes variables:

1. Probabilidad de que se encuentren n usuarios simultáneos ($Prob\{n = i\}$)
2. Tasa media de pérdida (*mean excess rate*)

La probabilidad de que se encuentren un número determinado de usuarios activos de forma simultánea, se obtiene en las simulaciones a partir de la recolección de datos de simulación, donde se van almacenando el periodo total en el que ha habido un número de usuarios activos determinado. En el caso del modelo analítico, esta probabilidad se obtiene utilizando la ecuación 5.15. La tasa media de pérdida se obtiene a partir del valor de la probabilidad anterior utilizando la ecuación 5.16 utilizando un valor de ejemplo de capacidad de servicio ($C = 250Mbps$).

En la tabla 5.2 se presentan los resultados extraídos de la ejecución de la simulación utilizando los parámetros y condiciones, descritos con anterioridad. En la tabla 5.3 se presentan las mismas variables y se añade el valor del error absoluto de la diferencia de probabilidades de que haya un número determinado de usuarios simultáneos en ambos modelos ($Prob\{n = i\}$).

A partir de los resultados de ambos modelos se puede obtener la tasa media de pérdida (ecuación 5.19), donde se aprecia que el error absoluto de ambos valores es de apenas el 0,2 %.

$$\begin{aligned} \text{Mean excess rate}\rfloor_{SIM} &= 1,423070185 \\ \text{Mean excess rate}\rfloor_{ANALITICO} &= 1,420863357 \end{aligned} \tag{5.19}$$

En la figura 5.6 se muestran las distribución de probabilidades de usuarios activos simultáneos en el sistema para el modelo de simulación y analítico respectivamente. Se puede apreciar como las distribuciones se solapan haciendo evidente el bajo error entre valores extraídos entre de los dos modelos.

$\mathbf{BW_{REQ}}$	Activos	Prob{n=i}	Perdidos	Excess rate
80	8	1,25E-07	0	0
90	9	5,36E-06	0	0
100	10	1,37E-05	0	0
110	11	4,31E-05	0	0
120	12	1,13E-04	0	0
130	13	3,39E-04	0	0
140	14	8,79E-04	0	0
150	15	1,97E-03	0	0
160	16	4,21E-03	0	0
170	17	8,37E-03	0	0
180	18	1,56E-02	0	0
190	19	2,67E-02	0	0
200	20	4,17E-02	0	0
210	21	5,92E-02	0	0
220	22	7,83E-02	0	0
230	23	9,52E-02	0	0
240	24	1,07E-01	0	0
250	25	1,12E-01	0	0
260	26	1,08E-01	1	1,08E-01
270	27	9,65E-02	2	1,93E-01
280	28	8,02E-02	3	2,41E-01
290	29	6,08E-02	4	2,43E-01
300	30	4,25E-02	5	2,13E-01
310	31	2,75E-02	6	1,65E-01
320	32	1,63E-02	7	1,14E-01
330	33	8,81E-03	8	7,05E-02
340	34	4,39E-03	9	3,95E-02
350	35	2,04E-03	10	2,04E-02
360	36	8,93E-04	11	9,82E-03
370	37	3,36E-04	12	4,04E-03
380	38	1,35E-04	13	1,75E-03
390	39	4,23E-05	14	5,92E-04
400	40	1,48E-05	15	2,22E-04
410	41	6,00E-06	16	9,60E-05
420	42	2,68E-07	17	4,56E-06
430	43	1,86E-07	18	3,35E-06

Tabla 5.2: Resultados de simulación para el escenario de validación

$\mathbf{BW_{REQ}}$	**Activos**	**Prob{n=i}**	**Excess rate**	$\mathbf{Error_{Prob}}$
80	8	4,55E-07	0	3,30E-07
90	9	2,13E-06	0	3,23E-06
100	10	8,75E-06	0	4,93E-06
110	11	3,19E-05	0	1,12E-05
120	12	1,04E-04	0	9,34E-06
130	13	3,05E-04	0	3,45E-05
140	14	8,08E-04	0	7,14E-05
150	15	1,94E-03	0	2,57E-05
160	16	4,26E-03	0	5,23E-05
170	17	8,55E-03	0	1,84E-04
180	18	1,57E-02	0	9,05E-05
190	19	2,66E-02	0	1,66E-04
200	20	4,13E-02	0	4,18E-04
210	21	5,91E-02	0	3,09E-05
220	22	7,82E-02	0	1,79E-04
230	23	9,54E-02	0	2,04E-04
240	24	1,08E-01	0	4,07E-04
250	25	1,12E-01	0	3,47E-04
260	26	1,08E-01	1,08E-01	6,97E-04
270	27	9,65E-02	1,93E-01	5,60E-05
280	28	7,95E-02	2,38E-01	7,22E-04
290	29	6,05E-02	2,42E-01	3,43E-04
300	30	4,24E-02	2,12E-01	7,67E-05
310	31	2,75E-02	1,65E-01	2,61E-05
320	32	1,63E-02	1,14E-01	3,77E-05
330	33	8,94E-03	7,15E-02	1,28E-04
340	34	4,48E-03	4,04E-02	9,45E-05
350	35	2,06E-03	2,06E-02	1,96E-05
360	36	8,59E-04	9,45E-03	3,41E-05
370	37	3,26E-04	3,91E-03	1,05E-05
380	38	1,12E-04	1,45E-03	2,30E-05
390	39	3,45E-05	4,83E-04	7,79E-06
400	40	9,51E-06	1,43E-04	5,26E-06
410	41	2,33E-06	3,72E-05	3,68E-06
420	42	5,00E-07	8,50E-06	2,32E-07
430	43	9,33E-08	1,68E-06	9,31E-08

Tabla 5.3: Resultados de modelo analítico para el escenario de validación

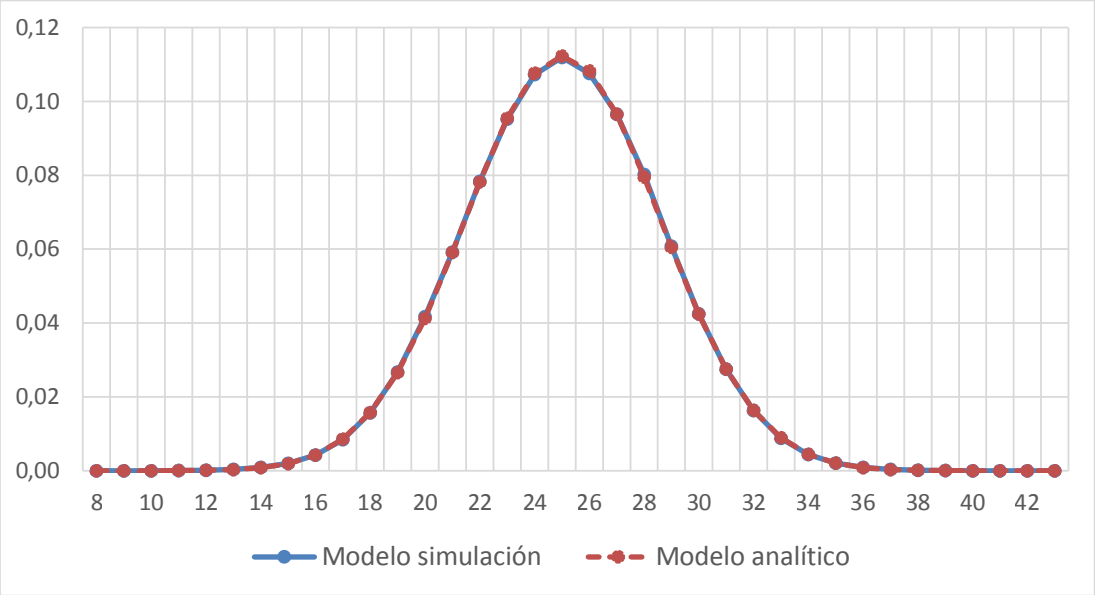

Figura 5.6: Distribuciones de probabilidad de usuarios activos simultáneos

5.3.4. Resumen y alcance de validación

En esta sección se presenta la validación entre un modelo analítico de un sistema de colas *burst scale buffering* frente a un modelo de simulación al que se le han impuesto un conjunto de condiciones.

Es importante resaltar que el modelo de simulación descrito en la sección 5.2.2 aborda un problema de recurso compartido para el cual se le puede aplicar un modelo analítico. Por esta razón, la validación descrita en esta sección tiene un alcance limitado a los siguientes objetivos:

- Confirmar la correcta implementación de la herramienta de simulación
- Validar parcialmente el modelo de simulación

Gracias a esta validación se puede afirmar que la herramienta de simulación desarrollada maneja y analiza con éxito los eventos discretos de múltiples fuentes de tráfico, así como la generación de eventos de actividad e inactividad. También se puede confirmar que la extracción de resultados de rendimiento (GoS se realizan correctamente ya que éstos dependen íntimamente de las algunas de las variables comparadas (como por ejemplo, la probabilidad de usuarios simultáneos).

No obstante, el alcance de esta validación se encuentra justamente relacionada con las condiciones impuestas en el modelo de simulación y que no pueden ser comparadas con el modelo analítico, como es el caso, de la superposición de fuentes de tráfico de

diferentes aplicaciones de Internet y los esquemas de asignación de ancho de banda entre aplicaciones y usuarios.

5.4. Caso de estudio 1: rendimiento en red de acceso del año 2012

En esta sección se aplica la metodología de estimación de demanda de tráfico y dimensionado de red, descrita a lo largo del capítulo 4, a un caso de estudio que representa una red de acceso con usuarios de Internet caracterizados a partir de datos estadísticos del año 2012. En este caso de estudio se hace uso del modelo de simulación descrito en la sección 5.2 que ha sido implementado en la herramienta de simulación descrita anteriormente.

5.4.1. Descripción de escenario de simulación

A continuación se describe el escenario utilizado como entrada en la herramienta de simulación, siguiendo como guía los parámetros de entrada definidos en la sección 5.2.4.2 de este mismo capítulo.

5.4.1.1. Parámetros de red de acceso

La red de acceso se modela mediante la definición de los siguientes parámetros de entrada:

- Número de suscriptores de la red de acceso
- Capacidades de los suscriptores (hogares)
- Número de usuarios de Internet por suscriptor

Número de suscriptores de red de acceso. Como se ha descrito con anterioridad en la sección 4.3.1, este parámetro de entrada varía en función del enlace que este siendo analizado con la herramienta de simulación. La razón reside en que la capacidad de suscriptores que tienen los agregadores de tráfico, como por ejemplo un DSLAM, varía en función de la arquitectura de red de acceso y el nivel de agregación en el que se encuentre el enlace analizado. Por ejemplo, los DSLAMs pueden tener capacidad de suscriptores que que varían entre unos pocos suscriptores (o dispositivos) soportados hasta decenas de miles de suscriptores soportados [Agilent Technologies, 2006].

Debido a que la aplicación a los casos de estudio suponen un ejemplo de uso de la metodología, se utiliza un valor de 1000 suscriptores de red de acceso. Este valor es lo suficientemente grande para ilustrar la flexibilidad de la herramienta y el impacto de la demanda de tráfico que tienen en la red de acceso.

Segmento de velocidad	**Líneas residenciales**	**Capacidad**	Proporción
$< 2Mbps$	196.435	2 Mbps	2,14 %
$\geq 2Mbps < 10Mbps$	2.984.077	6 Mbps	32,43 %
$\geq 10Mbps < 30Mbps$	5.017.669	20 Mbps	54,54 %
$\geq 30Mbps < 50Mbps$	587.776	30 Mbps	6,39 %
$\geq 50Mbps$	414.738	50 Mbps	4,51 %

Tabla 5.4: Capacidades de suscriptores de líneas residenciales [CNMC, 2012]

Capacidades de los suscriptores. Las capacidades de los suscriptores se puede definir como las velocidades que tienen contratadas los usuarios de Internet en sus hogares. Esta información puede ser extraída de los datos anuales de la CNMC de los apartados dedicados a las comunicaciones fijas y líneas de banda ancha.

En la tabla 5.4 se muestra el número de líneas de banda ancha fija correspondiente al segmento residencial [CNMC, 2012]. Además, también se muestran las velocidades representativas consideras, etiquetadas como *Capacidades* y su correspondiente proporción. Estas proporciones representan la probabilidad de que un suscriptor disponga de una capacidad u otra, por lo que éstas constituyen el parámetro de entrada para la herramienta.

Número de usuarios por suscriptor. Si se parte de la suposición que en la gran mayoría de los casos un suscriptor de red de acceso representa una vivienda u hogar, donde habitan un conjunto de posibles usuarios de Internet, se puede extraer este parámetro de entrada de la herramienta a partir de los datos estadísticos proporcionados por el INE.

En la tabla 5.5 se muestran los datos extraídos de [INE, 2012] y que indican el número de viviendas que disponen acceso a Internet con banda ancha en función del número de miembros que habitan en el hogar. A partir de estos datos, se pueden extraer las proporciones de número de usuarios que se encuentran en viviendas con acceso a Internet y que son equivalentes a los suscriptores de red de acceso. Estas proporciones extraídas conforman la distribución de número de usuarios por suscriptor de red de acceso y son otro parámetro de entrada de la simulación.

5.4.1.2. Parámetros de tráfico de red

El tráfico de red se define en la herramienta a partir de la definición de los siguientes parámetros de entrada:

- Conjunto de aplicaciones clasificadas como *background* o *foreground*
- Modelo de fuente de tráfico ON/OFF para cada aplicación

Vivienda	**Personas**	**Viviendas con BA**	**Personas**	Proporción
1 miembro	1	1.199.502	1.199.502	11,55 %
2 miembros	2	2.503.324	5.006.649	24,10 %
3 miembros	3	2.975.772	8.927.316	28,65 %
4 miembros	4	2.801.084	11.204.337	26,97 %
5 miembros	5	904.752	4.523.762	8,71 %
Total	2,97	10.387.659	30.861.565	100 %

Tabla 5.5: Distribución de personas por vivienda con banda ancha [INE, 2012]

A continuación, se enumeran las aplicaciones de Internet consideradas para este caso de estudio, definiendo para cada una su tipo:

1. Compartición de ficheros: *background*
2. Navegación web: *foreground*
3. Video sobre Internet: *foreground*
4. Juegos en red: *foreground*

Los modelos de fuente de tráfico ON/OFF de las aplicaciones de Internet, también considerados parámetros de entrada de la simulación, se encuentran implementados en la propia herramienta. Se utilizan los modelos seleccionados y descritos en la sección 5.2.3 de este mismo capítulo.

5.4.1.3. Parámetros de perfiles de usuario y aplicaciones

Los parámetros de entrada correspondientes al modelo de perfiles de usuario y aplicaciones que se definen en este caso de uso son los siguientes:

- Probabilidades de pertenencia a perfil de usuario
- Probabilidades y duraciones de conexión para cada perfil de usuario
- Matriz de probabilidades de uso de aplicaciones para cada perfil de usuario
- Porcentaje de concurrencia de usuarios de Internet

Probabilidades de pertenencia a perfil de usuario En la sección 5.4.1.1 se ha calculado el número de usuarios por suscriptor, pero no se ha tenido en cuenta la posibilidad de que existan individuos con acceso a Internet y que sin embargo no utilizan la red, es decir, que sean parte del segmento de *No-Usuarios*.

La probabilidad de pertenencia al perfil de *No-Usuarios* con acceso a Internet, p_0, se calcula a partir la ecuación 5.20. Las probabilidades mostradas en la ecuación se

p_i	Perfil de usuario	Probabilidad
p_0	No-Usuario con Internet	0,1286
p_1	Esporádicos	0,3115
p_2	Instrumentales	0,1795
p_3	Sociales	0,2655
p_4	Avanzados	0,2435

Tabla 5.6: Probabilidades de pertenencia a perfil de usuario

Variable	Esporádicos	Instrumentales	Sociales	Avanzados
pc_i	0,5057	0,8779	0,8986	0,9142
tc_i	88,1828	109,3279	130,7760	163,5114

Tabla 5.7: Probabilidades y tiempos de conexión de perfiles de usuario

extraen de la sección 3.6, siendo la proporción de *No-Usuarios* con acceso de un 7,9 % y la proporción de *Usuarios de Internet* de 53 %.

$$p_0 = \frac{\text{Prob}\{\text{No-usuarios con acceso}\}}{\text{Prob}\{\text{Usuarios Internet}\} + \text{Prob}\{\text{No-usuarios con acceso}\}} \tag{5.20}$$

A partir de los resultados del capítulo 3, descritos en la sección 3.6, se extraen las probabilidades de pertenencia a un perfil de usuarios de Internet. La tabla 5.6 muestra las probabilidades para cada perfil de usuario utilizadas como parámetros de entrada de la herramienta de simulación.

Probabilidades y duraciones de conexión Las probabilidades y duraciones de conexión de los perfiles de usuario de Internet se extraen a partir de la tabla 3.7 del capítulo 3, con los datos que caracterizan las conexiones desde el hogar. Los parámetros de entrada utilizados se muestran en la tabla 5.7, siendo pc_i la probabilidad de conexión y tc_i la duración de conexión de un perfil de usuario i.

Matriz de probabilidades de uso de aplicaciones A partir de los resultados derivados de la caracterización de usuario de Internet del capítulo 3 de esta tesis doctoral, se extraen un conjunto de indicadores de uso de aplicaciones para cada perfil de usuario de Internet identificado. No obstante, estos indicadores no sirven como parámetros de entrada de la herramienta, pues ésta espera una matriz que indique la probabilidad de uso de las aplicaciones para cada uno de los perfiles de usuario de Internet.

La tabla 5.8 muestra los indicadores de uso de las actividades en la red para cada uno de los perfiles de usuario de Internet. Estos valores también fueron representados

$\mathbf{V_{i,j}}$	**Actividad**	**Esporád.**	**Instrument.**	**Sociales**	**Avanzados**
$V_{1,j}$	Correo electrónico	2,55	4,91	4,78	5,19
$V_{2,j}$	Mensajería instantánea	1,31	2,18	3,63	4,00
$V_{3,j}$	Compart. de archivos	0,22	0,46	0,65	1,03
$V_{4,j}$	Telefonía sobre Internet	0,16	0,34	0,51	0,89
$V_{5,j}$	Juegos en red	0,20	0,24	0,65	0,78
$V_{6,j}$	Redes sociales	0,43	0,38	5,24	5,46
$V_{7,j}$	Blogs y foros	0,12	0,27	0,48	0,78
$V_{8,j}$	Lectura de Información	0,49	4,88	0,57	5,43
$V_{9,j}$	Programas y series de TV	0,24	0,61	0,87	1,74
$V_{10,j}$	Escuchar música	0,46	1,06	1,73	2,46
$V_{11,j}$	Podcasts	0,04	0,13	0,10	0,40
$V_{12,j}$	Banca Electrónica	0,25	0,87	0,36	1,05
$V_{13,j}$	Compra de productos/serv.	0,06	0,20	0,12	0,34

Tabla 5.8: Indicadores de uso de actividades para perfiles de usuario (veces por semana)

$\mathbf{U_{i,j}}$	**Aplicación de Internet**	**Esporád.**	**Instrument.**	**Sociales**	**Avanzados**
$U_{j,1}$	Compartición de ficheros	0,22	0,46	0,65	1,03
$U_{j,2}$	Video sobre Internet	0,65	2,12	3,24	5,45
$U_{j,3}$	Navegación web	2,01	9,90	10,15	19,58
$U_{j,4}$	Juegos en red	0,20	0,24	0,65	0,78

Tabla 5.9: Indicadores de uso de aplicaciones para perfiles de usuario (veces por semana)

en el capítulo 3 en la figura 3.10.

En el capítulo 4 se describe cómo se definen los indicadores de uso de aplicaciones de Internet a partir de la ecuación 4.10 (ver figura 4.11). En la ecuación 5.21 se muestran los valores utilizados para cada una de las aplicaciones de Internet del modelo.

$$\begin{aligned} U_{\text{File Sharing}}(P = P_j) &= U_{j,1} = V_{3,j} \\ U_{\text{Video}}(P = P_j) &= U_{j,2} = 0,3 \cdot V_{6,j} + 0,2 \cdot V_{8,j} + V_{9,j} + 0,4 \cdot V_{10,j} \\ U_{\text{Web}}(P = P_j) &= U_{j,3} = (V_{6,j} + V_{7,j} + V_{8,j} + V_{12,j} + V_{13,j}) \cdot 1,5 \\ U_{\text{Gaming}}(P = P_j) &= U_{j,4} = V_{5,j} \end{aligned} \tag{5.21}$$

La tabla 5.9 muestra los valores obtenidos para los indicadores de uso de aplicaciones después de aplicar la ecuación 5.21.

A continuación se extraen de estos indicadores de uso de aplicación de Internet las probabilidades de que un usuario de un perfil utilice cada aplicación. Para ello, se ha de tener en cuenta el modelo de tráfico de red descrito en la sección 4.3.2, ya que algunas aplicaciones serán de primer (*foreground*) o segundo plano (*background*).

Para el caso de las aplicaciones *background*, se define la probabilidad directamente a partir del indicador de uso pero cambiando la escala temporal de referencia de *veces por semana* a *veces al día*. En este caso de estudio, la única aplicación de este tipo es la

Aplicación de Internet	Esporádicos	Instrumentales	Sociales	Avanzados
Compartición de ficheros	0,0309	0,0659	0,0922	0,1475
Video sobre Internet	0,2276	0,1732	0,2309	0,2112
Navegación web	0,7039	0,8069	0,7227	0,7584
Juegos en red	0,0685	0,0199	0,0464	0,0304

Tabla 5.10: Matriz de probabilidades de uso de aplicaciones para perfiles de usuario

compartición de ficheros. Este tipo de aplicaciones tienden a estar activas durante largos periodos de tiempo. Por esta razón, para esta simulación se parte de la hipótesis que la probabilidad de que un usuario utilice la aplicación de a lo largo del día, es similar a la probabilidad de que utilice esa misma aplicación durante el periodo de máxima demanda de tráfico de la red de acceso. La probabilidad de uso de esta aplicación para cada uno de los perfiles se muestra en la primera fila de la tabla 5.10.

Para el caso de las aplicaciones *foreground*, sólo una de ellas se encontrará activa, por lo que se normalizan los indicadores de uso de aplicaciones de Internet con el fin de extraer una proporción (ecuación 5.22) que pueda utilizarse como una probabilidad. A partir d la segunda fila de la tabla 5.10 se muestran las probabilidades para las aplicaciones *foreground* consideradas.

$$Prob_{j,i} = \frac{U_{j,i}}{\sum_{i=2}^{4} U_{j,i}}, i \in [2,4] \tag{5.22}$$

La tabla 5.10 representa la matriz de probabilidades de uso de aplicaciones para cada uno de los perfiles de usuario de Internet, la cual es uno de los parámetros de entrada en la herramienta de simulación.

Concurrencia de usuarios de Internet Este parámetro de entrada suele ser utilizado para caracterizar la hora de máxima demanda de tráfico en la red de acceso, es decir, el equivalente de la *hora cargada* de redes de telefonía para redes de comunicaciones de datos.

El porcentaje de concurrencia utilizado habitualmente para caracterizar el periodo de máxima carga de demanda de tráfico y su impacto en el dimensionado de redes de acceso, suele encontrarse cerca del 20 % [Clark et al., 1999]. No obstante, debido a que las redes de acceso cada vez tienen que dar soporte a la demanda de aplicaciones con más requisitos de red, en este caso de estudio se define un valor de concurrencia de usuarios del 30 %.

5.4.1.4. Parámetros de simulación

Por último queda por definir los parámetros de simulación utilizados en la herramienta para la aplicación de este caso de uso:

- Tiempo de simulación (T_{SIM}) de 7400 segundos (2 horas y 200 segundos).
- Tamaño de la muestra ($S_{muestra}$) de 7400 muestras.
- Tiempo de inicio de fuente de tráfico (T_{START}) de 50 segundos.
- Tiempo de inicio de análisis de resultados ($T_{ANALYSIS}$) de 200 segundos.

El tiempo de simulación total analizado es de 2 horas ($T_{SIM} - T_{ANALYSIS}$) con una frecuencia de muestreo igual a 1 muestra por segundo. Debido al diseño del modelo y herramienta de simulación (sección 5.2.2, la frecuencia de muestreo no afecta a la calidad de los resultados finales de la simulación (este parámetro sólo tiene repercusión en el tamaño de algunas variables intermedias guardadas por la simulación). Se ha comprobado de forma heurística que el tiempo de inicio de fuente de tráfico (T_{START}) y el tiempo de inicio de análisis de resultados ($T_{ANALYSIS}$) son adecuados para haber alcanzado un estado estacionario de actividad de las fuentes de tráfico.

5.4.2. Resultados de simulación

A continuación se presentan los resultados de la aplicación del método para la estimación de demanda de tráfico para el caso de uso que representa una red de acceso en el año 2012. Los resultados se extraen a partir de la ejecución de la herramienta de simulación con los parámetros de entrada anteriores.

De forma análoga a la notación utilizada a lo largo del capítulo, las gráficas que se muestran a continuación utilizan índices para representar a los usuarios y aplicaciones de Internet respectivamente. Los perfiles de usuario de Internet se encuentran enumerados de la siguiente forma: *Usuarios Esporádicos* (1), *Usuarios Instrumentales* (2), *Usuarios Sociales* (3) y *Usuarios Avanzados* (4). De la misma forma, las aplicaciones de Internet también se encuentran enumeradas: *Compartición de ficheros* (1), *Video sobre Internet* (2), *Navegación web* (3) y *Juegos en red* (4).

El escenario de simulación de 1000 suscriptores de red (hogares) con un 30 % de concurrencia, hace que hayan 885 usuarios de Internet activos, con un media de capacidad de suscriptor de 16,91 Mbps.

En la figura 5.7 se muestran 3 gráficas con el fin de caracterizar al usuario de Internet de la ejecución del escenario de simulación. La primera gráfica muestra las capacidades de los usuarios de Internet que disponen en sus hogares expresadas en Mbps. La segunda y la tercera gráfica muestran los diferentes perfiles de usuario y aplicaciones de Internet que se encuentran activos en el escenario.

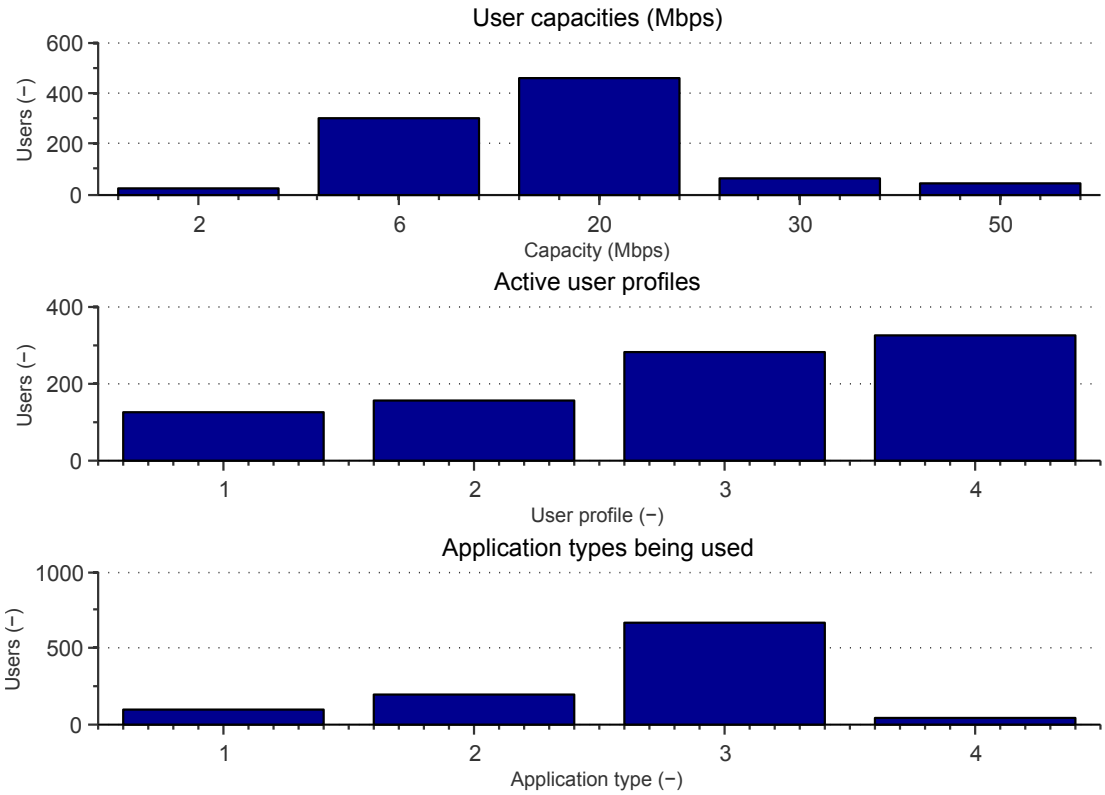

Figura 5.7: Resultados caso de estudio 1: usuarios de Internet en el escenario de simulación

Algunos de los resultados intermedios obtenidos, después ejecutar la simulación durante 18 horas y 12 minutos, son los siguientes:

- El número medio de usuarios activos simultáneos ha sido de 155
- El número de aplicaciones de Internet simultáneas ha sido de 170
- El factor de actividad promedio por usuario de Internet es de 17 %

En la figura 5.8 se muestran las distribuciones de usuarios activos y aplicaciones de Internet simultáneos. Se puede apreciar que los usuarios activos simultáneos oscilan entre los 130 y 180, mientras que las aplicaciones entre los 140 y 200.

Por último, en la figura 5.9 se muestra el rendimiento de la red de acceso para el escenario de simulación en términos de GoS. En la gráfica se aprecia que, por ejemplo, para una red de acceso proporcione un rendimiento en función del GoS requiere poder soportar un ancho de banda agregado determinado.

A continuación se extraen algunos valores de GoS en función del ancho de banda requerido en el enlace de la red de acceso, BW_{REQ}:

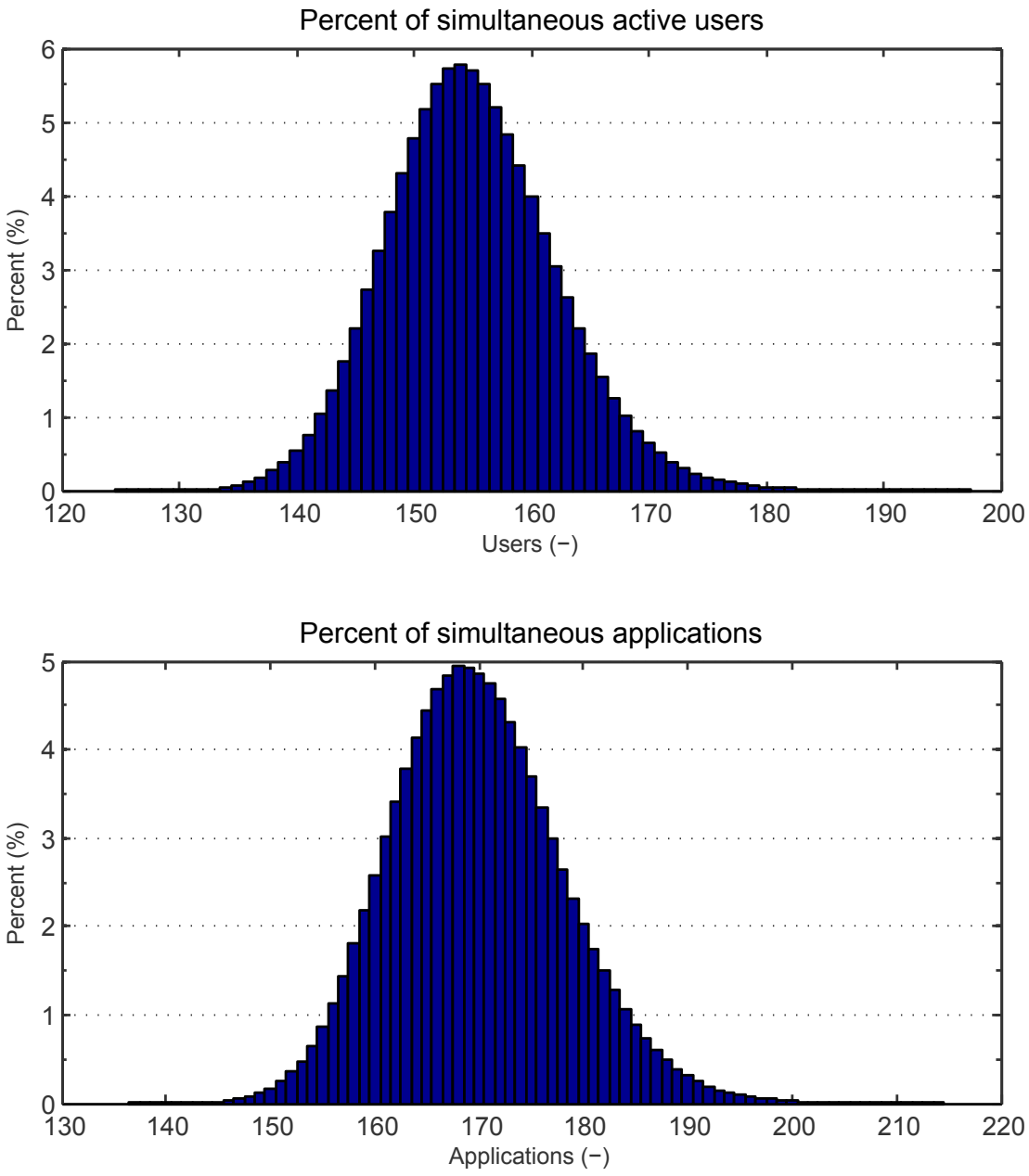

Figura 5.8: Resultados caso de estudio 1: usuarios y aplicaciones de Internet simultáneas

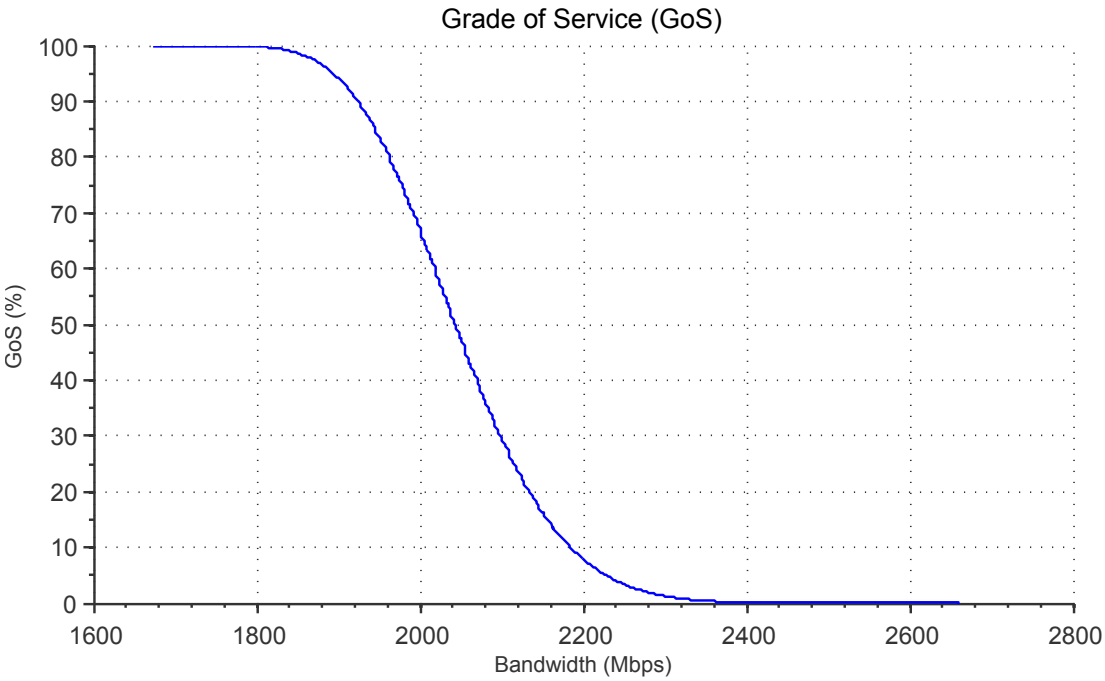

Figura 5.9: Resultados caso de estudio 1: rendimiento (GoS) de red de acceso

- $BW_{REQ} = 2185 Mbps$, se obtiene un $GoS = 10\,\%$
- $BW_{REQ} = 2230 Mbps$, se obtiene un $GoS = 5\,\%$
- $BW_{REQ} = 2312 Mbps$, se obtiene un $GoS = 1\,\%$
- $BW_{REQ} = 2416 Mbps$, se obtiene un $GoS = 0,1\,\%$
- $BW_{REQ} = 2502 Mbps$, se obtiene un $GoS = 0,01\,\%$
- $BW_{REQ} = 2575 Mbps$, se obtiene un $GoS = 0,001\,\%$

A partir del análisis del GoS se concluye que un operador puede aumentar considerablemente el rendimiento de la red de acceso en base a dimensionar correctamente la red de acceso. Se aprecia como el GoS de la red de acceso aumenta drásticamente con apenas aumentar un poco la capacidad del enlace que agrega el tráfico de los usuarios de Internet. Por ejemplo, para pasar de un GoS del 1 % al 0,01 % tan solo se requiere aumentar la capacidad cerca de un 8 %.

5.5. Caso de estudio 2: pronóstico de rendimiento en red de acceso

En esta sección se presenta la aplicación de la metodología de estimación de demanda de tráfico, descrita a lo largo del capítulo 4, a un caso de estudio que representa la misma red de acceso que en el caso de estudio anterior (sección 5.4), pero variando los parámetros de entrada relacionados con la caracterización de usuarios de Internet. Estos nuevos parámetros de entrada se extraen del pronóstico de la evolución de tipología de usuarios de Internet para el año 2017, descrito en la sección 3.7.2 del capítulo 3 de esta tesis.

El principal objetivo de este caso de estudio consiste en analizar el rendimiento de una red de acceso si se producen cambios en los hábitos y costumbres de consumo de aplicaciones por parte de los usuarios de Internet.

5.5.1. Descripción de escenario de simulación

En este estudio se definen los mismos parámetros de entrada que fueron utilizados anteriormente para el primer caso de estudio (sección 5.4.1), a excepción de las siguientes variables:

- Probabilidades de pertenencia a perfil de usuario
- Concurrencia de usuarios de Internet

En relación a los parámetros de red de acceso, se utiliza el mismo número de suscriptores (1000), y las mismas distribuciones de capacidades y de número de usuarios por suscriptor. Los parámetros de tráfico de red también se definen igual que en el caso de estudio anterior. Ambos conjuntos de parámetros de entrada no han sido modificados, ya que se pretende que la red de acceso del escenario de simulación sea la misma y poder así analizar el impacto de un cambio en la tipología.

Los parámetros de entrada de perfiles de usuario y aplicaciones se modifican para representar una evolución en la tipología de usuarios de Internet. Los principales cambios se realizan en las probabilidades de pertenencia a usuarios. Estos valores, mostrados en la tabla 5.11, han sido extraídos del pronóstico de la evolución de la tipología de usuarios de Internet, realizada en la sección 3.7.2 del capítulo 3. La probabilidad de ser *No-Usuario* con acceso a Internet ha sido definida en un 10 %. Este valor concuerda con que cada vez la proporción de *Usuarios de Internet* respecto a *No-Usuarios* ha ido aumentando. De la misma forma, la porción de *No-Usuario* con acceso a Internet desde sus hogares, también ha experimentado una reducción paulatina lo largo de los últimos años.

p_i	Perfil de usuario	Probabilidad
p_0	No-Usuario con Internet	0,10
p_1	Esporádicos	0,23
p_2	Instrumentales	0,12
p_3	Sociales	0,35
p_4	Avanzados	0,30

Tabla 5.11: Probabilidades de pertenencia a perfil de usuario en tipología pronosticada

Por último, se ha definido una concurrencia de usuarios de Internet del 40 %. Como se ha comentado con anterioridad, este parámetro de entrada caracteriza la hora de máxima demanda de tráfico de la red. Se considera el caso peor de que en los próximos años, la concurrencia aumente debido al incremento de la adopción tecnológica por parte de la población, constituyendo el consumo de aplicaciones de Internet una actividad cada vez más cotidiana.

El resto de parámetros de entrada de perfiles de usuario y aplicaciones de Internet se corresponden con los del caso de estudio anterior. La razón se encuentra en que la tipología de Internet ha sido extraída justamente a partir de los patrones de comportamiento y consumo de aplicaciones que se encuentran definidas en estas variables y que son las que caracterizan a los perfiles de usuario.

Los parámetros de simulación también se corresponden con los del caso de estudio anterior, es decir, se realiza una simulación de 2 horas y 200, y se comienza a recolectar los resultados pasados 200 segundos.

5.5.2. Resultados de simulación

A continuación se presentan los resultados de la aplicación del método de estimación de demanda al caso de estudio de una red de acceso caracterizada con el pronóstico de la evolución de los perfiles de usuarios de Internet. La notación utilizada en las figuras para identificar a los usuarios y aplicaciones de Internet, es la utilizada en el caso de estudio anterior (sección 5.4).

El escenario de simulación del caso de estudio ejecutado para 1000 suscriptores de red y un 40 % de concurrencia de usuarios, contiene 1179 usuarios de Internet activos con una media de 17,17 Mbps de capacidad suscriptor.

En la figura 5.10 se muestran 3 gráficas que caracterizan el escenario de la simulación desde el punto de vista del usuario de Internet. A partir de estos resultados, se aprecia la evolución de la tipología de usuarios de Internet hacia perfiles más intensivos, como los *Usuarios Sociales* o *Avanzados*. De la misma manera, este cambio en las proporciones de perfiles de usuario tiene una consecuencia directa en el aumento del resto de aplicaciones de Internet.

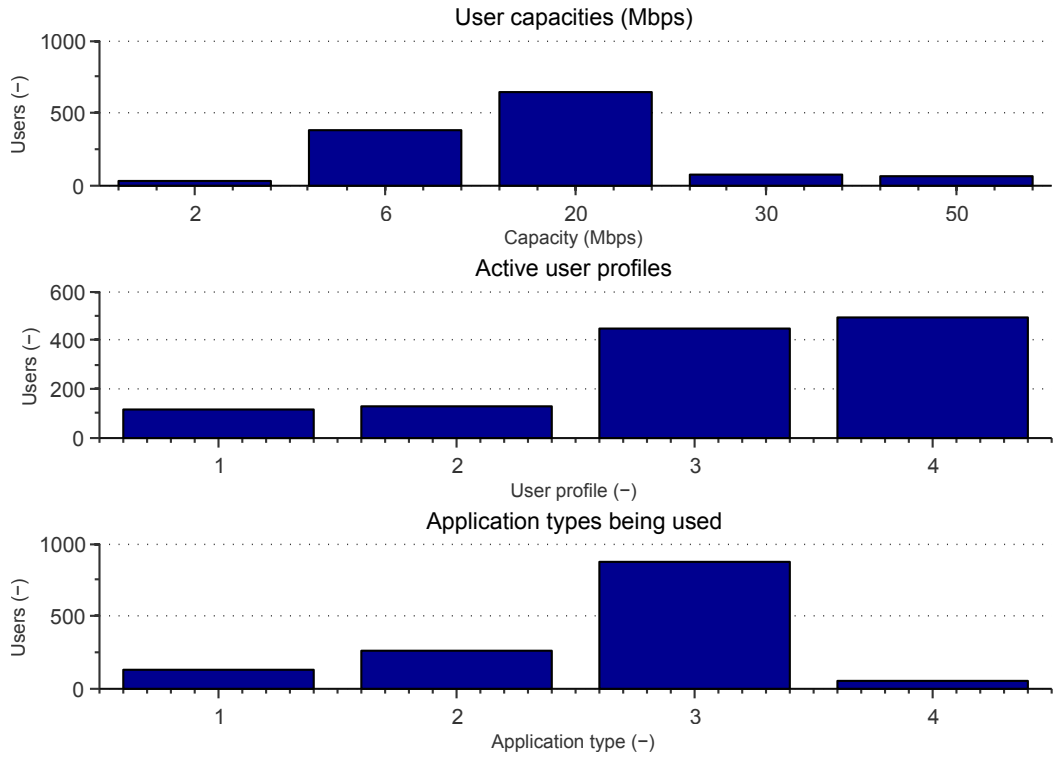

Figura 5.10: Resultados caso de estudio 2: usuarios de Internet en el escenario de simulación

Después de una simulación de 32 horas y 12 minutos, se extraen los siguientes resultados intermedios:

- El número medio de usuarios activos simultáneos ha sido de 217
- El número de aplicaciones de Internet simultáneas ha sido de 241
- El factor de actividad promedio por usuario de Internet es de 18 %

Estos resultados son coherentes con los obtenidos para el caso de estudio anterior, pues al haber más usuarios de Internet activos, también hay más usuarios y aplicaciones simultáneas. El factor de actividad ha aumentado de forma muy moderada, aunque era esperado debido a que hay más usuarios de los perfiles con mayor intensidad y variedad de consumo de aplicaciones de Internet.

En la figura 5.11 se muestran las distribuciones de usuarios activos y aplicaciones simultaneas. Se aprecia que existe los usuarios simultáneos oscilan entre 180 y 280 usuarios, y las aplicaciones simultaneas entre 210 y 290.

En la figura 5.12 se muestra el GoS en función del ancho de banda requerido en la red de acceso.

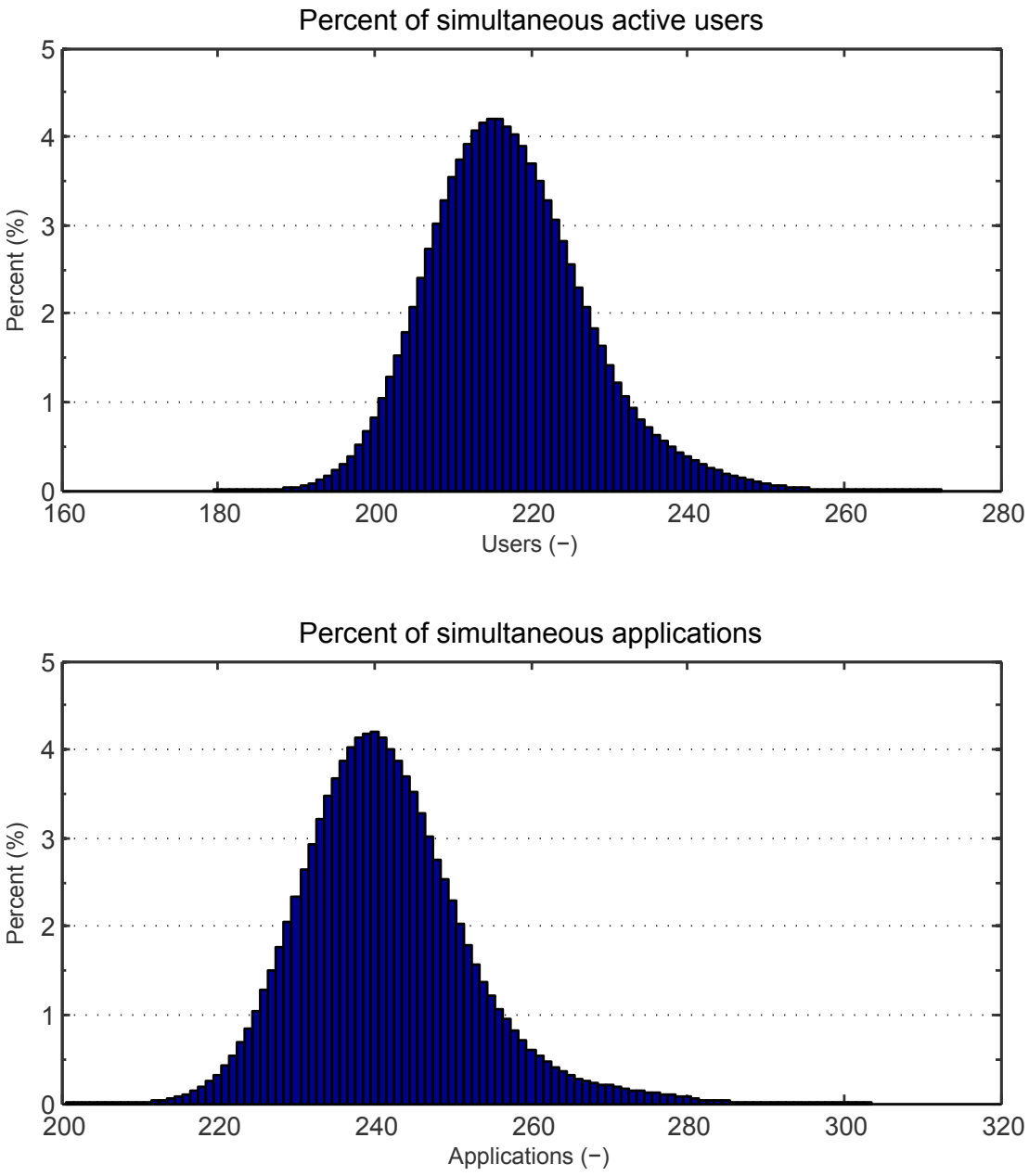

Figura 5.11: Resultados caso de estudio 1: usuarios y aplicaciones de Internet simultáneas

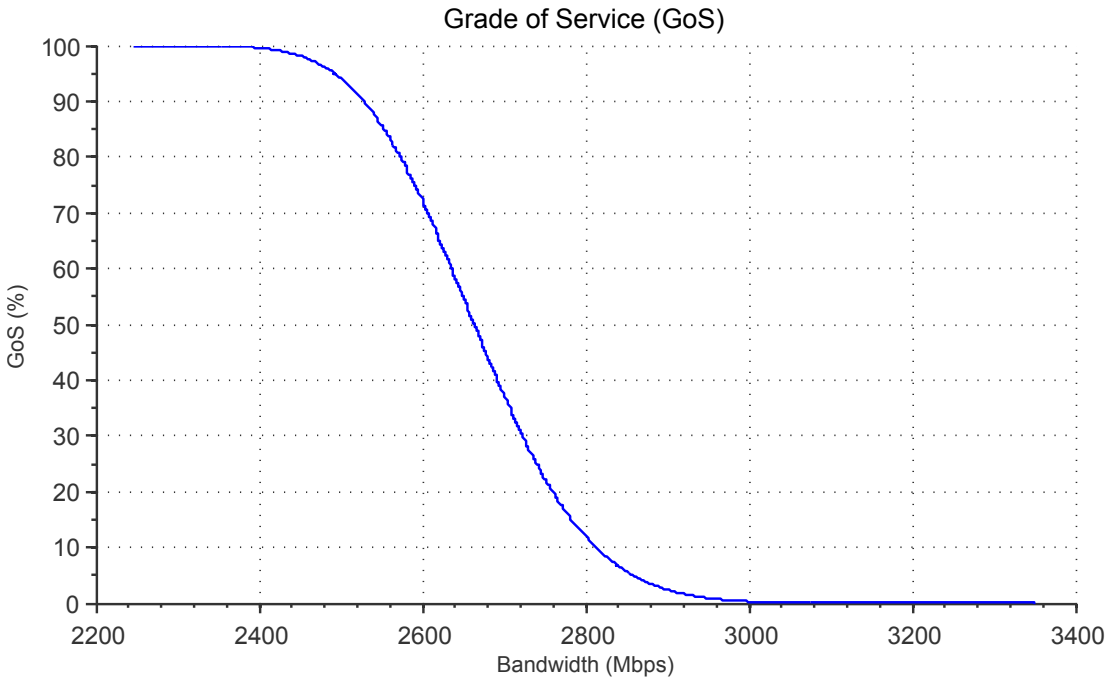

Figura 5.12: Resultados caso de estudio 2: rendimiento (GoS) de red de acceso

De forma análoga con el caso de estudio anterior, se extraen algunos valores de GoS en función del ancho de banda requerido en el enlace de la red de acceso, BW_{REQ}:

- $BW_{REQ} = 2814 Mbps$, se obtiene un $GoS = 10\,\%$
- $BW_{REQ} = 2860 Mbps$, se obtiene un $GoS = 5\,\%$
- $BW_{REQ} = 2948 Mbps$, se obtiene un $GoS = 1\,\%$
- $BW_{REQ} = 3048 Mbps$, se obtiene un $GoS = 0,1\,\%$
- $BW_{REQ} = 3136 Mbps$, se obtiene un $GoS = 0,01\,\%$
- $BW_{REQ} = 3212 Mbps$, se obtiene un $GoS = 0,001\,\%$

El análisis del GoS sobre este caso de estudio arroja las mismas conclusiones que para el caso de estudio anterior. Con una ligera mejora en la capacidad de los enlaces de la red de acceso se obtiene una mejora significativa del GoS ofrecido a los suscriptores de la red, o lo que es lo mismo, a los usuarios que acceden a Internet a través de la red de acceso. Por ejemplo, si una red de acceso con un GoS del 1 % aumenta un 9 % la capacidad del enlace de agregación de tráfico, podría aumentar el GoS de forma drástica hasta un 0,001 %.

Como se podía esperar, los anchos de banda requeridos en la red de acceso se han visto aumentados significativamente al introducir el pronóstico de evolución de tipología de usuarios de Internet. Por ejemplo, una red de acceso dimensionada para el primer caso de estudio con un GoS de 0,001 %, requeriría un ancho de banda de 2576 Mbps. Si se diesen las condiciones del segundo caso de estudio, es decir, la tipología de usuarios evoluciona y aumenta la concurrencia de usuarios a un 10 %, este ancho de banda de 2576 Mbps se corresponde con un GoS de 79 %. Este ejemplo, pone de manifiesto el gran impacto que sufren las redes de acceso a partir de cambios en la caracterización de los usuarios de Internet.

En definitiva, la aplicación del método de estimación de demanda de tráfico puede utilizarse para estimar el rendimiento en una red de acceso, así como, para establecer reglas de dimensionado de las mismas. También podría ser utilizado para analizar el impacto o la provisión necesaria en las redes de acceso ante cambios de los parámetros de los modelos que conforman el método de estimación, y que pueden ser traducidos a parámetros de entrada del modelo de simulación.

5.6. Conclusiones

En este capítulo se ha descrito el proceso de aplicación del método de estimación de demanda de tráfico, definido en el capítulo 4, a casos de estudio.

En primer lugar, el capítulo desarrolla un modelo de simulación en el que se define cómo se ha de aplicar el método de estimación de demanda mediante el uso de simulaciones basadas en eventos discretos. Además, se describe qué tipo de fuentes de tráfico se van a modelar, realizando una selección de los modelos de fuentes de tipo ON/OFF más adecuados para cada una de las aplicaciones de Internet consideradas en el método de estimación de demanda. Una vez definido el modelo de simulación, se describe la herramienta desarrollada durante esta tesis doctoral, que permite realizar simulaciones siguiendo este modelo.

En segundo lugar, se presenta una validación del modelo y herramienta de simulación mediante el uso de un escenario basado en múltiples fuentes homogéneas. Esta hipótesis ha permitido validar el modelo de simulación frente a un modelo analítico, obteniendo unos resultados satisfactorios.

El capítulo concluye con la aplicación del método a dos casos de estudio mediante el uso de la herramienta de simulación desarrollada. En el primer caso de estudio presenta una red de acceso, donde los perfiles de usuarios de Internet se encuentran caracterizados a partir de datos estadísticos correspondientes al año 2012. El segundo caso de estudio se basa en la misma red de acceso pero, en este caso, los usuarios de Internet se encuentran caracterizados por un pronóstico de evolución de los perfiles de usuario y un aumento de la concurrencia de usuarios de un 10 %. Tanto la caracterización de perfiles de usuario de

Internet, como el pronóstico de evolución de los perfiles, son obtenidas de los resultados del capítulo 3 de esta tesis doctoral.

La aplicación a casos de estudio demuestra la utilidad de la metodología de estimación de demanda de tráfico, permitiendo analizar y dimensionar las redes de acceso a Internet en función de una métrica de rendimiento. Esta aplicación también aporta una visión del impacto en las redes de acceso de los cambios en los hábitos y comportamiento de consumo por parte de diferentes grupos de usuarios de Internet, destacando la importancia de la identificación y caracterización de las tipologías de usuarios de Internet.

Capítulo 6

Conclusiones y líneas de trabajo futuras

En este capítulo se extraen los principales conclusiones de esta tesis doctoral a partir del análisis de los objetivos marcados en la sección 1.1. Posteriormente, se describe el marco de trabajo en el que se ha desarrollado esta tesis doctoral y se presenta un plan de explotación de resultados. Por último, se describe un conjunto de líneas de trabajo futuras con las que continuar la labor investigadora llevada a cabo durante el desarrollo de esta tesis doctoral.

6.1. Análisis de los objetivos

A continuación se analizan los objetivos descritos en la sección 1.1 y se resaltan las principales conclusiones extraídas.

6.1.1. Caracterización de usuarios de Internet en España

En el capítulo 3 se ha definido un marco metodológico robusto y conciso basado en KDPs para la identificación y caracterización de usuarios de Internet. A lo largo del capítulo, se han descrito y analizado cada una de las fases que conforman la metodología para asegurar la validez y calidad de los resultados extraídos.

A partir de la aplicación de esta metodología para el caso concreto de los usuarios residenciales de Internet en España, se ha identificado un conjunto de perfiles de usuarios que conforman la tipología de usuarios de Internet. Para esta identificación se han considerado múltiples variables específicas que indican hábitos de consumo de aplicaciones de Internet. Además, se ha realizado una caracterización de estos perfiles de usuario en función de las actividades que realizan en la red y otras variables socio-demográficas.

La aplicación de la metodología a los datos correspondientes al año 2012, ha proporcionado la identificación de *Usuarios de Internet* (53 %) y *No-Usuarios* (47 %). Los *No-Usuarios* representan aquellos individuos que no utilizan Internet o no tienen acceso a la red desde sus hogares.

La tipología de *Usuarios de Internet* se encuentra formada por los siguientes perfiles de usuario:

- *Esporádicos* (16 %): usuarios que utilizan la red de forma ocasional o infrecuente y que sobretodo utilizan servicios orientados a la comunicación (correo electrónico, mensajería instantánea, etc.).
- *Instrumentales* (10 %): usuarios que usan Internet como un instrumento, por lo que se caracterizan por el uso intensivo de servicios orientados a objetivos (búsqueda de información, banca electrónica, compras online, etc.).
- *Sociales* (14 %): usuarios caracterizados por un uso muy intensivo de las redes sociales en comparación con el uso que le dan a otros tipos de servicios. Muchos de ellos usan servicios orientados al entretenimiento (visionado de series y televisión, juegos en red, etc.)
- *Avanzados* (13 %): usuarios con mayor intensidad en frecuencia de uso de todos los servicios. Tienen un patrón de consumo variado e intenso. Son los que más utilizan servicios más específicos y complejos (compartición de archivos, visionado de series y televisión, etc.)

Esta metodología también ha sido aplicada a los datos correspondientes a años anteriores, con la intención de poder realizar un pronóstico de la evolución de la tipología de usuarios residenciales en España para los próximos años. A partir del análisis de los resultados, se espera que los perfiles de *Usuarios Esporádicos* e *Instrumentales* decrezcan un 7 % y un 14 % respectivamente, mientras que los *Usuarios Sociales* y *Avanzados* aumenten un 19 % y un 17 %.

Tanto la caracterización de usuarios de Internet realizada como el pronóstico de la evolución de la tipología, han sido utilizados como parámetros de entrada para la aplicación a casos de estudio de la metodología de estimación de demanda de tráfico y dimensionado de redes de acceso.

6.1.2. Propuesta de un método de estimación de demanda de tráfico y dimensionado de red de acceso

En el capítulo 4 se ha presentado y definido una metodología para la estimación de demanda de tráfico de Internet con el fin de posibilitar el análisis del rendimiento de la red de acceso subyacente.

Para describir el problema que se intenta abordar, se introduce y define el modelo teórico en el que se basa el método de estimación de demanda de tráfico y dimensionado de red de acceso. Además, se define la métrica de red GoS, para cuantificar el rendimiento de la red bajo unas condiciones determinadas.

El método presentado se basa en 3 modelos que caracterizan y definen los diferentes componentes necesarios para el análisis de demanda de tráfico en una red de acceso:

- Modelo de red de acceso: definición formal y caracterización de las entidades que componen el escenario de red de acceso. Este modelo considera diferentes niveles de agregación debido al acceso simultáneo de suscriptores de red (hogares), de usuarios y de aplicaciones de Internet.

- Modelo de tráfico de red: descripción y definición matemática de la superposición de demanda de tráfico correspondiente a los diferentes niveles de agregación considerados. Este modelo constituye el marco a partir del cual se definen los modelos matemáticos de fuente de tráfico que han de ser utilizados para estimar la demanda de tráfico de cada aplicación de Internet.

- Modelo de perfiles de usuario y aplicaciones: caracterización del uso de aplicaciones de Internet por parte de los perfiles de usuario. Este modelo define el uso de aplicaciones para cada tipo de usuario teniendo en cuenta la actividad de conexión y los patrones de consumo de aplicaciones de los distintos perfiles.

Este método, basado en los 3 modelos anteriores, constituye el marco metodológico para la estimación de demanda de tráfico de Internet y el posterior análisis de rendimiento de la red de acceso.

6.1.3. Desarrollo de un modelo y herramienta de simulación

Antes de proceder a la aplicación de los casos de estudio, el capítulo 5 presenta un modelo y herramienta de simulación que posibilita la aplicación del método de estimación de demanda de tráfico de Internet y dimensionado de red de acceso, descrito en el capítulo 4.

Este modelo de simulación tiene como objetivo definir un marco para la estimación la demanda de tráfico y rendimiento de la red de acceso mediante el uso de herramientas de simulación discreta de eventos. El modelo tiene la finalidad de describir formalmente los siguientes aspectos:

- Definir el tipo de fuentes de tráfico que se van a modelar considerando diferentes niveles de agregación de tráfico (suscriptor, usuario y aplicación).

- Marcar los objetivos de la simulación, teniendo como principal parámetro de salida la métrica de rendimiento de red utilizada en esta tesis, el GoS.

- Definir un proceso de simulación a partir del cual se puedan alcanzar los obtenidos anteriores.
- Seleccionar los modelos de fuente de tráfico tipo ON/OFF para cada aplicación de Internet considerada en los escenarios de simulación.

Haciendo uso del modelo de simulación anterior, se desarrolla una herramienta de simulación, desarrollada en *Matlab*, para aplicar el método de estimación de demanda de tráfico y dimensionado de red de acceso.

A partir de un escenario sencillo con fuentes de tráfico homogéneas y sin niveles de agregación, se realiza una validación del modelo y herramienta de simulación, al comparar los resultados de la simulación con los calculados a partir de un modelo analítico. Los resultados han confirmado el correcto desarrollo e implementación del modelo de simulación en la herramienta.

6.1.4. Aplicación del método a casos de estudio

En el capítulo 5 se aplica el método a dos casos de estudio mediante el uso de la anterior herramienta de simulación para analizar el rendimiento de una misma red de acceso y considerando los mismos modelos de fuente de tráfico de aplicaciones:

- Red de acceso en el año 2012: la caracterización de los perfiles de usuario y el uso que hacen de las aplicaciones de Internet se basa en los resultados obtenidos a lo largo del capítulo 3.
- Red de acceso en el año 2017: la caracterización de los perfiles de usuario y el uso que hacen de las aplicaciones de Internet se basa en el pronóstico de la evolución de la tipología de usuarios de Internet realizada en el capítulo 3.

A partir de estos casos de uso se aporta una aplicación del método de estimación de demanda de tráfico de usuario y se analiza el impacto que tiene un cambio en la tipología de usuarios de Internet en el rendimiento (GoS) y dimensionado de las redes de acceso.

6.2. Difusión de resultados

El trabajo realizado durante esta tesis doctoral, así como sus contribuciones, han sido de gran utilidad para el desarrollo del proyecto de investigación “ *VideoXperience*: Mejora Efectiva de la Experiencia de Usuario en la Nueva Era de Servicios Digitales mediante la Provisión de nuevas Tecnologías de Supercompresión en Streaming”. Este proyecto forma parte del subprograma INNPACTO del Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica 2008-2011, y ha sido financiado por el

Ministerio de Ciencia e Innovación, que se corresponde actualmente al Ministerio de Economía y Competitividad.

Los objetivos primordiales del proyecto *VideoXperience* son los siguientes:

- Caracterizar el dimensionamiento de Internet para poder ofrecer servicios de vídeo de alta calidad con una experiencia de usuario medible y similar a los actuales sistemas de TDT e IPTV desplegados por operadores.

- Cubrir el gap existente entre los resultados obtenidos con el primer objetivo y la capacidad de las redes actuales. Para ello se desarrollará un nuevo sistema de codificación de imagen y video capaz de satisfacer dicha experiencia de usuario en Internet sobre cualquier red de acceso fija o móvil. Esto reducirá el coste por byte, aumentará la capacidad de las redes existentes y mejorará la experiencia de usuario.

En el contexto de este proyecto de investigación se ha realizado difusión de resultados mediante las siguientes publicaciones:

- P. De la Cruz Ramos, M. Cao Cueto, R. Pérez Leal, F. González Vidal, **Estimation of Perceived Quality in Convergent Services**, Proc. of the Seventh IARIA International Conference on Digital Telecommunications (ICDT 2012), pp.88-95. Chamonix, France. Abril-Mayo 2012.

- Jose Javier García Aranda, Marina González Casquete, Mario Cao Cueto, Joaquín Navarro Salmerón, Francisco González Vidal. **Logarithmical hopping encoding: a low computational complexity algorithm for image compression.** Aceptado para publicación en IET Image Processing Journal.

6.3. Plan de explotación de resultados

A continuación se describe el plan de explotación de los resultados obtenidos a lo largo de la elaboración de esta tesis doctoral.

6.3.1. Identificación de conocimientos extraídos

En primer lugar se identifican y valoran los conocimientos científicos y técnicos extraídos que pueden ser explotados, como por ejemplo, en forma de productos o procesos:

- Metodología basada en KDPs para la caracterización y extracción de perfiles de usuarios de Internet a partir de fuentes de información estadísticas.

- Método de estimación de demanda de tráfico de Internet y dimensionado de redes de acceso.
- Herramienta de simulación para analizar rendimiento y dimensionado de redes de acceso.

Estas contribuciones conforman una innovación tecnológica desde el punto de vista que suponen un nexo de unión entre dos áreas de conocimiento muy diferentes, el estudio de fuentes de información estadísticas y los métodos matemáticos y formales para el análisis de rendimiento y dimensionado de redes de acceso. Las principales ventajas o mejoras tecnológicas respecto a los procesos disponibles en la actualidad son las siguientes:

- Se considera la existencia de una tipología de usuarios de Internet, es decir, un conjunto de perfiles de usuarios que difieren en patrones de uso y consumo de servicios.
- Las metodologías descritas en esta tesis se encuentran desarrolladas siguiendo un esquema altamente modular, por lo que sus componentes pueden ser modificados sencillamente para adaptarse a cambios futuros de la red, como por ejemplo la inclusión de nuevos perfiles de usuario o servicios de Internet.
- El método de estimación de demanda de tráfico y dimensionado de red de acceso permite de forma sencilla modificar los parámetros de entrada, lo cual posibilita analizar el efecto que producen cambios en los patrones de comportamiento de los usuarios o los servicios de Internet.

La herramienta de simulación, que implementa el método de estimación de demanda y dimensionado de red de acceso, ha sido desarrollada como un prototipo con el principal objetivo de demostrar la validez y aplicabilidad del método. Por esta razón, la explotación de resultados podría requerir la revisión del desarrollo de la herramienta con el fin de que pudiera estar lista para producción.

Una de las principales barreras potenciales para la explotación o transferencia de los resultados reside en la visibilidad que se doten a los mismos. Por este motivo, se ha de llevar un ambicioso plan de difusión de resultados con objeto de aumentar significativamente la visibilidad de los potenciales productos, procesos o servicios que puedan derivarse de los resultados.

6.3.2. Identificación de participantes, usuarios y mercados potenciales

Esta tesis doctoral ha sido realizada en el marco de diversos proyectos de investigación llevados por la Universidad Politécnica de Madrid, entre los que destaca el anteriormente

mencionado *VideoXperience*. De ente los participantes del proyecto, destaca *Alcatel-Lucent*, una multinacional francesa proveedora de equipamiento, software y servicios a proveedores de servicios de telecomunicaciones. Debido a que gran parte de su negocio se centra en la venta de equipamiento para redes móviles y redes de datos, así como para la distribución de video y televisión, esta empresa podría constituir un usuario potencialmente interesado en los resultados de esta tesis doctoral. Por ejemplo, podrían proveer equipamiento de red en función del estudio de la demanda de usuarios de Internet y su correspondiente efecto en el dimensionado de la red de acceso.

Los usuarios potenciales que más interesados podrían estar en los resultados de esta tesis doctoral se encuentran en el sector de las telecomunicaciones, como por ejemplo, ISPs, proveedores de servicios y contenidos de Internet, etc. Principalmente se podrían beneficiar de los usos citados para el análisis de rendimiento y dimensionamiento de las redes de acceso.

Otros usuarios potenciales pueden estar interesados en utilizar los resultados de esta tesis para el análisis del impacto en las redes de acceso de un avance tecnológico que afecte al ancho de banda. Por ejemplo, el uso de un servicio o algoritmo que pueda aumentar o disminuir el ancho de banda requerido por los usuarios de Internet. Este podría ser el caso de usuarios que pertenezcan a comunidades científicas que requieran de una herramienta que les permita analizar el impacto, positivo o negativo, de la inclusión de un avance tecnológico.

Además, otro mercado potencial se encuentra en el ámbito del marketing y la publicidad. Esta tesis doctoral realiza una caracterización socio-demográfica de los diferentes tipos de usuarios de Internet, la cual puede ser utilizada para conocer cómo se encuentra segmentado el mercado. Esta información podría ser muy valiosa para todos aquellos interesados en realizar campañas de publicidad u ofertas a un conjunto de usuarios específicos. Esta técnica es conocida como *ad targeting* y puede ser de especial interés para operadores y proveedores de servicios y contenidos de Internet.

6.3.3. Plan de difusión de resultados

Como se ha identificado anteriormente, la principal barrera asociada a la explotación de resultados se encuentra en la visibilidad de los mismos por aquellos usuarios y mercados potenciales. Para abordar este problema se ha elaborado el siguiente plan de difusión de resultados con el objetivo de hacer más visible el trabajo realizado durante esta tesis doctoral.

1. Publicación de la memoria de esta tesis doctoral en el Archivo Digital de la UPM [UPM, 2015]. Los documentos de este archivo son indexados por las principales buscadores y fuentes de información científicas.

2. Publicación de la herramienta de simulación de eventos discretos en un repositorio público de *GitHub* [GitHub, 2015]. El código fuente puede ser descargado, utilizado e incluso modificado, siempre y cuando se cumpla con la licencia.

3. Publicaciones de artículos científicos en revistas internacionales con las principales contribuciones de esta tesis doctoral. Por ejemplo, artículos sobre la metodología basada en KDPs para caracterizar y extraer perfiles de usuarios de Internet, y sobre el método de estimación de demanda de tráfico de Internet a partir de la caracterización de perfiles de usuario.

En relación a la protección intelectual de los resultados, se ha optado por distribuir la herramienta de simulación bajo una licencia Apache (versión 2). De esta forma, una parte de los resultados de esta tesis doctoral se encuentran visibles y disponibles para cualquier usuario o empresa interesada, siempre y cuando los trabajos derivados cumplan con los requisitos de la licencia.

6.4. Líneas de trabajo futuras

En esta sección se presentan algunas líneas de investigación, que han sido identificadas durante el desarrollo de esta tesis doctoral y que pueden ser objeto de interés para continuar o complementar las contribuciones presentadas en esta memoria:

- Ampliación del estudio a usuarios de Internet que acceden a través de redes y dispositivos móviles: a pesar del auge de los *smartphones* y otros dispositivos móviles, las fuentes de información estadísticas, analizadas en el capítulo 2, aún no disponen de suficientes conjuntos de datos para la aplicación de la metodología basada en KDP para la caracterización de usuarios de Internet.

- A partir de la caracterización de usuarios de Internet "con movilidad", aplicación de metodología de estimación de demanda de tráfico a escenario de redes de acceso móviles: esta caracterización posibilitaría el análisis del impacto de la evolución en la tipología de usuarios de Internet en el futuro en el rendimiento y dimensionado para redes de acceso de tecnologías móviles.

- Ampliación de estudio de tipologías de usuarios de Internet utilizando datos estadísticos de otros años: la caracterización de usuarios de Internet requiere una constante vigilancia de los cambios en los comportamientos de usuarios de Internet. La aparición de nuevas aplicaciones o el aumento del consumo de algunas aplicaciones de Internet, pueden propiciar que las técnicas de minería de datos identifiquen nuevos conglomerados y patrones de comportamiento de consumo. Es posible que este fenómeno pueda darse en breve debido al constante aumento del consumo de aplicaciones de video sobre Internet.

- Mejora del modelo de tráfico para aplicaciones de compartición de ficheros: identificación o desarrollo de un modelo de fuente de tráfico ON/OFF representativo para aplicaciones de compartición de ficheros que considere algunas características propias, como por ejemplo, el número de ficheros o el tamaño de los contenidos descargados. También podría ser interesante implementar un modelo de fuente de tráfico ON/OFF para alguna aplicación representativa de este tipo de servicios.

- Mejora del modelo de tráfico para aplicaciones de video sobre Internet: identificación o desarrollo de un modelo de fuente de tráfico ON/OFF para la aplicación de video sobre Internet que tenga en cuenta nuevas técnicas de streaming o codificación. Esta línea de investigación se encuentra en línea con el auge de nuevas aplicaciones de video basadas en técnicas de streaming dinámico adaptativo, como por ejemplo, MPEG-DASH. También podría ser de especial interés, considerar otras técnicas de codificación de video que permitan reducir la demanda de tráfico de este tipo de aplicaciones.

- Análisis de impacto en el rendimiento de redes de acceso para una nueva técnica o algoritmo que permita reducir el ancho de banda demandado por una aplicación de Internet: dado una nueva técnica o algoritmo que permita reducir o modificar los patrones de tráfico de una aplicación de Internet, se puede analizar el grado de mejora de rendimiento experimentado en la red de acceso. Esto puede realizarse mediante la implementación de un modelo de fuente de tráfico que caracterice la mejora de la aplicación y el uso de la metodología de estimación de demanda y dimensionado, propuesta en esta tesis doctoral.

Bibliografía

[Abhari and Soraya, 2010] Abhari, A. and Soraya, M. (2010). Workload generation for youtube. *Multimedia Tools and Applications*, 46(1):91–118.

[Adtran, 2009] Adtran (2009). Defining broadband speeds: Deriving required capacity in access networks. Technical report, Adtran.

[Agilent Technologies, 2006] Agilent Technologies (2006). Understanding dslam and bras access devices. *Agilent Technologies white paper*.

[Aidouni et al., 2009] Aidouni, F., Latapy, M., and Magnien, C. (2009). Ten weeks in the life of an edonkey server. In *Parallel & Distributed Processing, 2009. IPDPS 2009. IEEE International Symposium on*, pages 1–5. IEEE.

[AIMC, 2013a] AIMC (2013a). Audiencia de Internet en el EGM. Technical report, Asociación para la Investigación de Medios de Comunicación.

[AIMC, 2013b] AIMC (2013b). Navegantes en la red - encuesta aimc a usuarios de internet. Technical report, Asociación para la Investigación de Medios de Comunicación.

[Ajzen and Fishbein, 1977] Ajzen, I. and Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological bulletin*, 84(5):888.

[Álvarez-Campana et al., 2009] Álvarez-Campana, M., Berrocal Colmenarejo, J., González Vidal, F., Pérez Leal, R., Román Martínez, I., and Vázquez Gallo, E. (2009). *Tecnologías de banda ancha y convergencia de redes*. Ministerio de Industria, Turismo y Comercio (España).

[AMETIC, 2011] AMETIC (2011). Informe anual 2011. Technical report, Asociación de Empresas de Electrónica, Tecnologías de la Información, Telecomunicaciones y Contenidos Digitales.

[AMETIC, 2012a] AMETIC (2012a). Always on. always connected. Technical report, Asociación de Empresas de Electrónica, Tecnologías de la Información, Telecomunicaciones y Contenidos Digitales.

[AMETIC, 2012b] AMETIC (2012b). Informe de la industria de contenidos digitales 2012. Technical report, Asociación de Empresas de Electrónica, Tecnologías de la Información, Telecomunicaciones y Contenidos Digitales.

[Anick et al., 1982] Anick, D., Mitra, D., and Sondhi, M. (1982). Stochastic theory of a data-handling system with multiple sources. *The Bell System Technical Journal*, 61(8):1871–1894.

[Armitage et al., 2006] Armitage, G., Claypool, M., and Branch, P. (2006). *Networking and online games: understanding and engineering multiplayer Internet games*. John Wiley & Sons.

[Asensio et al., 2008] Asensio, E., Ordua, J., and Morillo, P. (2008). Analyzing the network traffic requirements of multiplayer online games. In *Advanced Engineering Computing and Applications in Sciences, 2008. ADVCOMP'08. The Second International Conference on*, pages 229–234. IEEE.

[Baarsch and Celebi, 2012] Baarsch, J. and Celebi, M. E. (2012). Investigation of internal validity measures for k-means clustering. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1, pages 14–16.

[Baiocchi et al., 1991] Baiocchi, A., Melazzi, N. B., Listanti, M., Roveri, A., and Winkler, R. (1991). Loss performance analysis of an atm multiplexer loaded with high-speed on-off sources. *IEEE Journal on Selected Areas in Communications*, 9(3):388–393.

[Barakat et al., 2003] Barakat, C., Thiran, P., Iannaccone, G., Diot, C., and Owezarski, P. (2003). Modeling internet backbone traffic at the flow level. *IEEE Transactions on Signal Processing*, 51(8):2111–2124.

[Basher et al., 2008] Basher, N., Mahanti, A., Mahanti, A., Williamson, C., and Arlitt, M. (2008). A comparative analysis of web and peer-to-peer traffic. In *Proceedings of the 17th international conference on World Wide Web*, pages 287–296. ACM.

[Bellman, 1956] Bellman, R. (1956). Dynamic programming and lagrange multipliers. *Proceedings of the National Academy of Sciences of the United States of America*, 42(10):767.

[Berners-Lee et al., 1996] Berners-Lee, T., Fielding, R. T., and Nielsen, H. F. (1996). RFC 1945 – Hypertext Transfer Protocol – HTTP/1.0. `http://www.faqs.org/rfcs/rfc1945.html`.

[Bertsekas et al., 1992] Bertsekas, D. P., Gallager, R. G., and Humblet, P. (1992). *Data networks*, volume 2. Prentice-Hall International New Jersey.

[Borella, 2000] Borella, M. S. (2000). Source models of network game traffic. *Computer Communications*, 23(4):403–410.

[Brachman and Anand, 1996] Brachman, R. J. and Anand, T. (1996). The process of knowledge discovery in databases. In *Advances in knowledge discovery and data mining*, pages 37–57. American Association for Artificial Intelligence.

[Bradley and Fayyad, 1998] Bradley, P. S. and Fayyad, U. M. (1998). Refining initial points for k-means clustering. In *ICML*, volume 98, pages 91–99. Citeseer.

[Brandtzæg, 2010] Brandtzæg, P. B. (2010). Towards a unified media-user typology (mut): A meta-analysis and review of the research literature on media-user typologies. *Computers in Human Behavior*, 26(5):940–956.

[Brandtzæg et al., 2011] Brandtzæg, P. B., Heim, J., and Karahasanović, A. (2011). Understanding the new digital divide—a typology of internet users in europe. *International journal of human-computer studies*, 69(3):123–138.

[Brichet et al., 1996] Brichet, F., Roberts, J., Simonian, A., and Veitch, D. (1996). Heavy traffic analysis of a storage model with long range dependent on/off sources. *Queueing systems*, 23(1-4):197–215.

[Cabena et al., 1998] Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., and Zanasi, A. (1998). *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[CableLabs, 1996] CableLabs (1996). Specifications library - DOCSIS. URL: `http://www.cablelabs.com/specs/`. [Online: Febrero 2015].

[Cáceres et al., 1991] Cáceres, R., Danzig, P. B., Jamin, S., and Mitzel, D. J. (1991). Characteristics of wide-area tcp/ip conversations. In *Proceedings of the Conference on Communications Architecture &Amp; Protocols*, SIGCOMM '91, pages 101–112, New York, NY, USA. ACM.

[Caliński and Harabasz, 1974] Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

[Chen et al., 2005] Chen, K.-T., Huang, P., Huang, C.-Y., and Lei, C.-L. (2005). Game traffic analysis: An mmorpg perspective. In *Proceedings of the international workshop on Network and operating systems support for digital audio and video*, pages 19–24. ACM.

[Chen et al., 2008] Chen, Y., Wang, W., Fu, L., and Zhang, X. (2008). Traffic model for http video page. In *Communications and Networking in China, 2008. ChinaCom 2008. Third International Conference on*, pages 432–436. IEEE.

[Chiu et al., 2001] Chiu, T., Fang, D., Chen, J., Wang, Y., and Jeris, C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. In *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*, pages 263–268. ACM.

[Choi and Limb, 1999] Choi, H.-K. and Limb, J. O. (1999). A behavioral model of web traffic. In *Network Protocols, 1999.(ICNP'99) Proceedings. Seventh International Conference on*, pages 327–334. IEEE.

[Çiflikli et al., 2010] Çiflikli, C., Gezer, A., Özşahin, A. T., and Özkasap, Ö. (2010). Bittorrent packet traffic features over ipv6 and ipv4. *Simulation Modelling Practice and Theory*, 18(9):1214–1224.

[Cios et al., 2010] Cios, K. J., Pedrycz, W., Swiniarski, R. W., and Kurgan, L. A. (2010). *Data mining: a knowledge discovery approach*. Springer Publishing Company, Incorporated.

[CIS, 2012] CIS (2012). Barómetro de junio 2012.

[Cisco, 2013] Cisco (2013). The zettabyte era–trends and analysis. *Cisco white paper*.

[Cisco, 2014] Cisco (2014). Cisco visual networking index: forecast and methodology, 2013-2018. Technical report, Cisco.

[Clark et al., 1999] Clark, D., Lehr, W., and Liu, I. (1999). Provisioning for bursty internet traffic: Implications for industry and internet structure. In *Proc. MIT ITC Workshop on Internet Quality of Service*. Citeseer.

[CMT, 2011] CMT (2011). Informe económico sectorial. Technical report, Comisión del Mercado de las Telecomuncaciones.

[CNMC, 2012] CNMC (2012). Informe anual 2012. Technical report, Comisión Nacional de los Mercados y la Competencia.

[Costa et al., 2004] Costa, C. P., Cunha, I. S., Borges, A., Ramos, C. V., Rocha, M. M., Almeida, J. M., and Ribeiro-Neto, B. (2004). Analyzing client interactivity in streaming media. In *Proceedings of the 13th international conference on World Wide Web*, pages 534–543. ACM.

[Cunche et al., 2012] Cunche, M., Kaafar, M. A., Chen, J., Boreli, R., and Mahanti, A. (2012). Why are they hiding? study of an anonymous file sharing system. In *Satellite Telecommunications (ESTEL), 2012 IEEE First AESS European Conference on*, pages 1–6. IEEE.

[Dainotti et al., 2005] Dainotti, A., Pescape, A., and Ventre, G. (2005). A packet-level traffic model of starcraft. In *Hot Topics in Peer-to-Peer Systems, 2005. HOT-P2P 2005. Second International Workshop on*, pages 33–42. IEEE.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

[Deng, 1996] Deng, S. (1996). Empirical model of www document arrivals at access link. In *Communications, 1996. ICC'96, Conference Record, Converging Technologies for Tomorrow's Applications. 1996 IEEE International Conference on*, volume 3, pages 1797–1802. IEEE.

[Dziuban and Shirkey, 1974] Dziuban, C. D. and Shirkey, E. C. (1974). When is a correlation matrix appropriate for factor analysis? some decision rules. *Psychological Bulletin*, 81(6):358.

[Erman et al., 2005] Erman, D., Ilie, D., and Popescu, A. (2005). Bittorrent session characteristics and models. In *3rd International Conference HET-NETs'05*.

[Erman et al., 2006] Erman, D., Ilie, D., and Popescu, A. (2006). Bittorrent traffic characteristics. In *Computing in the Global Information Technology, 2006. ICCGI'06. International Multi-Conference on*, pages 42–42. IEEE.

[Ester et al., 1996] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

[Estivill-Castro, 2002] Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. *ACM SIGKDD Explorations Newsletter*, 4(1):65–75.

[ETSI, 2003a] ETSI (2003a). Access and terminals (at); second generation transmission systems for interactive cable television services - ip cable modems. ETSI Standard (ES) 202 488, European Telecommunications Standards Institute.

[ETSI, 2003b] ETSI (2003b). Digital broadband cable access to the public telecommunications network; ip multimedia time critical services;. Technical Specification (TS) 101 909, European Telecommunications Standards Institute.

[ETSI, 2009] ETSI (2009). Universal mobile telecommunications system (umts); technical specifications and technical reports for a utran-based 3gpp system (3gpp ts 21.101 version 8.0.0 release 8). Technical Specification (TS) 121 101, European Telecommunications Standards Institute.

[ETSI, 2011a] ETSI (2011a). Universal mobile telecommunications system (umts); technical specifications and technical reports for a utran-based 3gpp system (3gpp ts 21.101 version 10.0.0 release 10). Technical Specification (TS) 121 101, European Telecommunications Standards Institute.

[ETSI, 2011b] ETSI (2011b). Universal mobile telecommunications system (umts); technical specifications and technical reports for a utran-based 3gpp system (3gpp ts 21.101 version 9.0.1 release 9). Technical Specification (TS) 121 101, European Telecommunications Standards Institute.

[EUROSTAT and Seybert, 2012] EUROSTAT and Seybert, H. (2012). Internet use in households and by individuals in 2012. Technical report, Statistical Office of the European Communities.

[Fan-Bin Zeng, 2011] Fan-Bin Zeng, D. (2011). Impact factors model of internet adoption and use: taking the college students as an example. *Global Journal of Human-Social Science Research*, 11(7).

[Färber, 2002] Färber, J. (2002). Network game traffic modelling. In *Proceedings of the 1st workshop on Network and system support for games*, pages 53–57. ACM.

[Färber et al., 1999] Färber, J., Bodamer, S., and Charzinski, J. (1999). Statistical evaluation and modeling of internet dial-up traffic. In *Photonics East'99*, pages 112–121. International Society for Optics and Photonics.

[Fayyad et al., 1996a] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996a). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37.

[Fayyad et al., 1996b] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996b). Advances in knowledge discovery and data mining. American Association for Artificial Intelligence.

[Feknous et al., 2014] Feknous, M., Houdoin, T., Le Guyader, B., De Biasio, J., Gravey, A., and Torrijos Gijon, J. (2014). Internet traffic analysis: A case study from two major european operators. In *Computers and Communication (ISCC), 2014 IEEE Symposium on*, pages 1–7.

[Fielding et al., 1999] Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., and Berners-Lee, T. (1999). Rfc 2616, hypertext transfer protocol – http/1.1.

[Forconi et al., 2008] Forconi, S., Iazeolla, G., Kritzinger, P., and Pillegi, P. (2008). Modelling internet workloads for ieee 802.16. *University of Cape Town: South Africa, Rep. CS08-03-00*.

[Fred et al., 2001] Fred, S. B., Bonald, T., Proutiere, A., Régnié, G., and Roberts, J. W. (2001). Statistical bandwidth sharing: a study of congestion at flow level. In *ACM SIGCOMM Computer Communication Review*, volume 31, pages 111–122. ACM.

[Fundación Orange, 2012] Fundación Orange (2012). eespaña, informe anual 2012 sobre el desarrollo de la sociedad de la información en españa. Technical report, Fundación Orange.

[Fundación Telefónica, 2013] Fundación Telefónica (2013). La sociedad de la información en españa 2012. Technical report, Fundación Telefónica.

[García et al., 2007] García, R., Pañeda, X. G., García, V., Melendi, D., and Vilas, M. (2007). Statistical characterization of a real video on demand service: User behaviour and streaming-media workload analysis. *Simulation Modelling Practice and Theory*, 15(6):672–689.

[Gehlen et al., 2012] Gehlen, V., Finamore, A., Mellia, M., and Munafò, M. M. (2012). *Uncovering the big players of the web*. Springer.

[GitHub, 2015] GitHub (2015). Github repository: discrete-event simulation tool for access network bandwidth allocation. Disponible en `https://github.com/mariocao/bandwidth-sim`.

[Glaropoulos et al., 2014] Glaropoulos, I., Luna, A. V., Fodor, V., and Papadopouli, M. (2014). Closing the gap between traffic workload and channel occupancy models for 802.11 networks. *Ad Hoc Networks*, 21:60–83.

[INE, 2012] INE (2012). Encuesta sobre equipamiento y uso de tecnologías de la información y comunicación en los hogares. Nuevas tecnologías de la información y la comunicación.

[Greiner et al., 1999] Greiner, M., Jobmann, M., and Lipsky, L. (1999). The importance of power-tail distributions for modeling queueing systems. *Operations Research*, 47(2):313–326.

[Grimm and Schlüchtermann, 2008] Grimm, C. and Schlüchtermann, G. (2008). *IP-Traffic Theory and Performance*. Springer.

[Han et al., 2006] Han, J., Kamber, M., and Pei, J. (2006). *Data mining, southeast asia edition: Concepts and techniques*. Morgan kaufmann.

[Han et al., 2012] Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, third edition.

[Hartigan, 1975] Hartigan, J. A. (1975). Clustering algorithms.

[Hassan et al., 2005] Hassan, H., Garcia, J.-M., and Brun, O. (2005). Generic modeling of multimedia traffic sources.

[Hastie et al., 2005] Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85.

[Hatem et al., 1997] Hatem, J., Lipsky, L., and Fiorini, P. (1997). Comparison of buffer usage utilizing multiple servers in network systems with power-tail distributions. In *in Network Systems With Power-Tail Distributions. INFORMS97, Boston MA, 30 June-2*, page http://www.engr.ucon.

[Hawa et al., 2012] Hawa, M., Rahhal, J. S., and Abu-Al-Nadi, D. I. (2012). File size models for shared content over the bittorrent peer-to-peer network. *Peer-to-peer Networking and Applications*, 5(3):279–291.

[He et al., 2007] He, G., Hou, J., Chen, W.-P., and Hamada, T. (2007). One size does not fit all: a detailed analysis and modeling of p2p traffic. In *Global Telecommunications Conference, 2007. GLOBECOM'07. IEEE*, pages 393–398. IEEE.

[Heath et al., 1998] Heath, D., Resnick, S., and Samorodnitsky, G. (1998). Heavy tails and long range dependence in on/off processes and associated fluid models. *Mathematics of Operations Research*, 23(1):145–165.

[Höflich and Rössler, 2001] Höflich, J. R. and Rössler, P. (2001). Mobile schriftliche kommunikation oder: E-mail für das handy. *Medien & Kommunikationswissenschaft*, 49(4):437–461.

[Horrigan, 2009] Horrigan, J. (2009). The mobile difference. *Pew Internet & American Life Project*.

[Horrigan, 2007] Horrigan, J. B. (2007). A typology of information and communication technology users.

[Howard et al., 2001] Howard, P. E., Rainie, L., and Jones, S. (2001). Days and nights on the internet the impact of a diffusing technology. *American Behavioral Scientist*, 45(3):383–404.

[IEEE, 2002] IEEE (2002). Ieee standard for local and metropolitan area networks - part 16: Air interface for fixed broadband wireless access systems. IEEE Standard 802.16-2001, Institute of Electrical and Electronics Engineers.

[IEEE, 2004] IEEE (2004). Ieee standard for local and metropolitan area networks - part 16: Air interface for fixed broadband wireless access systems. IEEE Standard 802.16-2004, Institute of Electrical and Electronics Engineers.

[IEEE, 2006] IEEE (2006). Amendment for physical and medium access control layers for combined fixed and mobile operation in licensed bands. IEEE Standard 802.16e-2005, Institute of Electrical and Electronics Engineers.

[Im et al., 2011] Im, I., Hong, S., and Kang, M. S. (2011). An international comparison of technology adoption: Testing the utaut model. *Information & Management*, 48(1):1–8.

[ITU, 2012a] ITU (2012a). Measuring the information society 2012. Technical report, International Telecommunication Union.

[ITU, 2012b] ITU (2012b). Measuring the information society 2012. Technical report, Broadband commission of the International Telecommunication Union.

[ITU-T, 1999] ITU-T (1999). Asymmetric digital subscriber line (adsl) transceivers. Recommendation G.992.1, International Telecommunication Union.

[ITU-T, 2001] ITU-T (2001). Very high speed digital subscriber line transceivers (vdsl). Recommendation G.992.5, International Telecommunication Union.

[ITU-T, 2002] ITU-T (2002). Asymmetric digital subscriber line transceivers 2 (adsl2). Recommendation G.992.3, International Telecommunication Union.

[ITU-T, 2003a] ITU-T (2003a). Asymmetric digital subscriber line 2 transceivers (adsl2)-extended bandwidth adsl2 (adsl2plus). Recommendation G.992.5, International Telecommunication Union.

[ITU-T, 2003b] ITU-T (2003b). Gigabit-capable passive optical networks (gpon): General characteristics. Recommendation G.984.1, International Telecommunication Union.

[ITU-T, 2006] ITU-T (2006). Very high speed digital subscriber line transceivers 2 (vdsl2). Recommendation G.992.5, International Telecommunication Union.

[ITU-T, 2010] ITU-T (2010). 10-gigabit-capable passive optical networks (xg-pon): General requirements. Recommendation G.987.1, International Telecommunication Union.

[Jain, 2010] Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666.

[Jain and Routhier, 1986] Jain, R. and Routhier, S. (1986). Packet trains–measurements and a new model for computer network traffic. *IEEE Journal on Selected Areas in Communications*, 4(6):986–995.

[Johnson and Kulpa, 2007] Johnson, G. M. and Kulpa, A. (2007). Dimensions of online behavior: Toward a user typology. *CyberPsychology & Behavior*, 10(6):773–780.

[Katsaros et al., 2012] Katsaros, K. V., Xylomenos, G., and Polyzos, G. C. (2012). Globetraff: a traffic workload generator for the performance evaluation of future internet architectures. In *2012 5th International Conference on New Technologies, Mobility and Security (NTMS)*, pages 1–5. IEEE.

[Katz et al., 1999] Katz, E., Blumler, J. G., and Gurevitch, M. (1999). Utilization of mass communication by the individual. *Sources notable selections in mass media*, pages 51–59.

[Kihl et al., 2010] Kihl, M., Odling, P., Lagerstedt, C., and Aurelius, A. (2010). Traffic analysis and characterization of internet user behavior. In *Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2010 International Congress on*, pages 224–231. IEEE.

[Labovitz et al., 2011] Labovitz, C., Iekel-Johnson, S., McPherson, D., Oberheide, J., and Jahanian, F. (2011). Internet inter-domain traffic. *ACM SIGCOMM Computer Communication Review*, 41(4):75–86.

[Lai et al., 2014] Lai, P. K., Chow, K., Hui, L. C., and Yiu, S. (2014). Modeling the initial stage of a file sharing process on a bittorrent network. *Peer-to-Peer Networking and Applications*, 7(4):311–319.

[Lang and Armitage, 2003] Lang, T. and Armitage, G. (2003). A ns2 model for the xbox system link game halo. In *in Proc. Australian Telecommunications Networks and Applications Conference*.

[Lang et al., 2003] Lang, T., Armitage, G., Branch, P., and Choo, H.-Y. (2003). A synthetic traffic model for half-life. In *Australian Telecommunications Networks & Applications Conference*, volume 2003.

[Lang et al., 2004] Lang, T., Branch, P., and Armitage, G. (2004). A synthetic traffic model for quake3. In *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology*, pages 233–238. ACM.

[Lee and Gupta, 2007] Lee, J. J. and Gupta, M. (2007). A new traffic model for current user web browsing behavior. *Intel Corporation*.

[Leibowitz et al., 2002] Leibowitz, N., Bergman, A., Ben-Shaul, R., and Shavit, A. (2002). Are file swapping networks cacheable? characterizing p2p traffic. In *Proc. of the 7th Int. WWW Caching Workshop*.

[Li et al., 2013] Li, B., Wang, Z., Liu, J., and Zhu, W. (2013). Two decades of internet video streaming: A retrospective view. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 9(1s):33.

[Likert, 1932] Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

[Likhanov et al., 1995] Likhanov, N., Tsybakov, B., and Georganas, N. D. (1995). Analysis of an atm buffer with self-similar ("fractal") input traffic. In *INFOCOM'95. Fourteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Bringing Information to People. Proceedings. IEEE*, pages 985–992. IEEE.

[Limaye et al., 2008] Limaye, P. S., Glapa, M. J., El-Sayed, M. L., and Gagen, P. F. (2008). Impact of bandwidth demand growth on hfc networks. In *Telecommunications Network Strategy and Planning Symposium, 2008. Networks 2008. The 13th International*, pages 1–10. IEEE.

[Liu et al., 2008] Liu, Y., Guo, Y., and Liang, C. (2008). A survey on peer-to-peer video streaming systems. *Peer-to-peer Networking and Applications*, 1(1):18–28.

[Lloyd, 1982] Lloyd, S. (1982). Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, 28(2):129–137.

[Luo, 2005] Luo, S. (2005). *Creating models of internet background traffic suitable for use in evaluating network intrusion detection systems*. PhD thesis, University of Central Florida Orlando, Florida.

[Luo and Marin, 2005] Luo, S. and Marin, G. A. (2005). Realistic internet traffic simulation through mixture modeling and a case study. In *Proceedings of the 37th Conference on Winter Simulation*, WSC '05, pages 2408–2416. Winter Simulation Conference.

[MacQueen et al., 1967] MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

[Mah, 1997] Mah, B. A. (1997). An empirical model of http network traffic. In *INFOCOM'97. Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Driving the Information Revolution., Proceedings IEEE*, volume 2, pages 592–600. IEEE.

[Mahanti et al., 2012] Mahanti, A., Carlsson, N., Arlitt, M. F., and Williamson, C. (2012). Characterizing cyberlocker traffic flows. In *LCN*, pages 410–418.

[Maier et al., 2009] Maier, G., Feldmann, A., Paxson, V., and Allman, M. (2009). On dominant characteristics of residential broadband internet traffic. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 90–102. ACM.

[Mao and Jain, 1996] Mao, J. and Jain, A. K. (1996). A self-organizing network for hyperellipsoidal clustering (hec). *Neural Networks, IEEE Transactions on*, 7(1):16–29.

[Marbán et al., 2009] Marbán, Ó., Mariscal, G., and Segovia, J. (2009). A data mining & knowledge discovery process model. *Data Mining and Knowledge Discovery in Real Life Applications*, pages 1–17.

[McQuail, 2010] McQuail, D. (2010). *McQuail's mass communication theory*. Sage publications.

[Mikosch et al., 2002] Mikosch, T., Resnick, S., Rootzén, H., and Stegeman, A. (2002). Is network traffic approximated by stable lévy motion or fractional brownian motion? *Annals of Applied Probability*, 12(1):23–68.

[Milligan, 1980] Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, 45(3):325–342.

[Milligan, 1996] Milligan, G. W. (1996). *Clustering validation: results and implications for applied analyses*. World Scientific.

[Milligan and Cooper, 1985] Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179.

[Montagnier and Wirthmann, 2011] Montagnier, P. and Wirthmann, A. (2011). Digital divide: From computer access to online activities–a micro data analysis.

[OFCOM, 2008] OFCOM (2008). Social networking: A quantitative and qualitative research report into attitudes, behaviours and use.

[ONTSI, 2006] ONTSI (2006). Perfil sociodemográfico de los internautas. actividades realizadas en internet. iii trimestre 2003-iii trimestre 2005. las tic en los hogares españoles. Technical report, Observatorio Nacional de las Telecomunicaciones y Sociedad de la Información.

[ONTSI, 2013a] ONTSI (2013a). Las tic en los hogares españoles. datos de actitudes, usos, equipamiento y gasto tic correspondientes al tercer trimestre del 2012. Technical report, Observatorio Nacional de las Telecomunicaciones y Sociedad de la Información.

[ONTSI, 2013b] ONTSI (2013b). Perfil sociodemográfico de los internautas, análisis de datos ine 2012. Technical report, Observatorio Nacional de las Telecomunicaciones y Sociedad de la Información.

[Ortega Egea et al., 2006] Ortega Egea, J. M., Recio Menéndez, M., and Román González, M. V. (2006). Diffusion and usage patterns of internet services in the european union. *Information Research*, 12(2):15.

[Park and Willinger, 2000] Park, K. and Willinger, W. (2000). *Self-similar network traffic and performance evaluation*. Wiley Online Library.

[Paxson, 1994] Paxson, V. (1994). Empirically derived analytic models of wide-area tcp connections. *IEEE/ACM Trans. Netw.*, 2(4):316–336.

[Paxson and Floyd, 1995] Paxson, V. and Floyd, S. (1995). Wide area traffic: the failure of poisson modeling. *IEEE/ACM Transactions on Networking (ToN)*, 3(3):226–244.

[Peña et al., 1999] Peña, J. M., Lozano, J. A., and Larranaga, P. (1999). An empirical comparison of four initialization methods for the k-means algorithm. *Pattern recognition letters*, 20(10):1027–1040.

[Pesovic and Sharpe, 2012] Pesovic, A. and Sharpe, R. (2012). Is symmetrical bandwidth a myth or a must. *TechZine<http://www2. alcatel-lucent. com/techzine/is-symmetrical-bandwidth-a-myth-or-a-must/>(accessed 22.11. 2013.)*.

[Pitts and Schormans, 2001] Pitts, J. M. and Schormans, J. A. (2001). *Introduction to IP and ATM design and performance: with applications analysis software*. John Wiley & Sons, Inc.

[Postel and Reynolds, 1985] Postel, J. and Reynolds, J. (1985). Ietf rfc 959 file transfer protocol: Ftp.

[Pries et al., 2012] Pries, R., Magyari, Z., and Tran-Gia, P. (2012). An http web traffic model based on the top one million visited web pages. In *Next Generation Internet (NGI), 2012 8th EURO-NGI Conference on*, pages 133–139. IEEE.

[Rao et al., 2011] Rao, A., Legout, A., Lim, Y.-s., Towsley, D., Barakat, C., and Dabbous, W. (2011). Network characteristics of video streaming traffic. In *Proceedings of the Seventh COnference on emerging Networking EXperiments and Technologies*, page 25. ACM.

[Ratti et al., 2010] Ratti, S., Hariri, B., and Shirmohammadi, S. (2010). A survey of first-person shooter gaming traffic on the internet. *Internet Computing, IEEE*, 14(5):60–69.

[Reyes-Lecuona et al., 1999] Reyes-Lecuona, A., González-Parada, E., Casilari, E., Casasola, J., and Diaz-Estrella, A. (1999). A page-oriented www traffic model for wireless system simulations. In *Proceedings ITC*, volume 16, pages 1271–1280.

[Robert and Le Boudec, 1997] Robert, S. and Le Boudec, J.-Y. (1997). New models for pseudo self-similar traffic. *Performance Evaluation*, 30(1):57–68.

[Rogers, 2010] Rogers, E. M. (2010). *Diffusion of innovations*. Simon and Schuster.

[Schwartz, 1996] Schwartz, M. (1996). *Broadband integrated networks*, volume 19. Prentice Hall PTR New Jersey.

[Schwefel and Lipsky, 1999] Schwefel, H.-P. and Lipsky, L. (1999). Performance results for analytic models of traffic in telecommunication systems, based on multiple on-off sources with self-similar behavior. In *Elsevier Science B.V*, pages 55–66. Elsevier.

[Schwefel and Lipsky, 2001] Schwefel, H.-P. and Lipsky, L. (2001). Impact of aggregated, self-similar on/off traffic on delay in stationary queueing models (extended version). *Perform. Eval.*, 43(4):203–221.

[Selwyn et al., 2005] Selwyn, N., Gorard, S., and Furlong, J. (2005). Whose internet is it anyway? exploring adults'(non) use of the internet in everyday life. *European Journal of Communication*, 20(1):5–26.

[Shearer, 2000] Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22.

[Shih and Venkatesh, 2004] Shih, C.-F. and Venkatesh, A. (2004). Beyond adoption: Development and application of a use-diffusion model. *Journal of Marketing*, 68(1):59–72.

[Shroff and Schwartz, 1998] Shroff, N. B. and Schwartz, M. (1998). Improved loss calculations at an atm multiplexer. *IEEE/ACM Transactions on Networking (TON)*, 6(4):411–421.

[Sivogolovko, 2013] Sivogolovko, E. (2013). The influence of data quality on clustering outcomes. In *Databases and Information Systems VII: Selected Papers from the Tenth International Baltic Conference, DB&IS 2012*, volume 249, page 95. IOS Press.

[SPSS, 2011] SPSS (2011). SPSS 20.0 command syntax reference.

[Srinivasan et al., 2008] Srinivasan, R., Zhuang, J., Jalloul, L., Novak, R., and Park, J. (2008). Ieee 802.16 m evaluation methodology document (emd). *IEEE 802.16 Broadband Wireless Access Working Group*.

[Svoboda, 2008] Svoboda, P. (2008). *Measurement and modelling of Internet traffic over 2.5 and 3G cellular core networks*. PhD thesis.

[Tan et al., 2006] Tan, P.-N., Steinbach, M., Kumar, V., et al. (2006). *Introduction to data mining*, volume 1. Pearson Addison Wesley Boston.

[Tang et al., 2003] Tang, W., Fu, Y., Cherkasova, L., and Vahdat, A. (2003). Medisyn: A synthetic streaming media service workload generator. In *Proceedings of the 13th international workshop on Network and operating systems support for digital audio and video*, pages 12–21. ACM.

[Taqqu et al., 1997] Taqqu, M. S., Willinger, W., and Sherman, R. (1997). Proof of a fundamental result in self-similar traffic modeling. *ACM SIGCOMM Computer Communication Review*, 27(2):5–23.

[Tibshirani et al., 2001] Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

[UPM, 2015] UPM (2015). Archivo digital UPM. Disponible en `http://oa.upm.es`.

[Veloso et al., 2002] Veloso, E., Almeida, V., Meira, W., Bestavros, A., and Jin, S. (2002). A hierarchical characterization of a live streaming media workload. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*, pages 117–130. ACM.

[Venkatesh et al., 2003] Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS quarterly*, pages 425–478.

[Venkatesh et al., 2012] Venkatesh, V., Thong, J. Y., and Xu, X. (2012). Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS quarterly*, 36(1):157–178.

[Vicari, 1997] Vicari, N. (1997). *Measurement and modeling of WWW-sessions*. Inst. für Informatik.

[Vu-Brugier, 2009] Vu-Brugier, G. (2009). Analysis of the impact of early fiber access deployment on residential internet traffic. In *Teletraffic Congress, 2009. ITC 21 2009. 21st International*, pages 1–8. IEEE.

[Wamser et al., 2011] Wamser, F., Pries, R., Staehle, D., Heck, K., and Tran-Gia, P. (2011). Traffic characterization of a residential wireless internet access. *Telecommunication Systems*, 48(1-2):5–17.

[Wattimena, 2006] Wattimena, A. (2006). Performance modeling of interactive gaming. *Vrije Universiteit Amsterdam, Tech. Rep.*

[Willinger et al., 1998] Willinger, W., Paxson, V., and Taqqu, M. S. (1998). Self-similarity and heavy tails: Structural modeling of network traffic. *A practical guide to heavy tails: statistical techniques and applications*, 23:27–53.

[Willinger et al., 1997] Willinger, W., Taqqu, M. S., Sherman, R., and Wilson, D. V. (1997). Self-similarity through high-variability: statistical analysis of ethernet lan traffic at the source level. *IEEE/ACM Transactions on Networking*, 5(1):71–86.

[Xu et al., 2014] Xu, K., Shen, M., Cui, Y., Ye, M., and Zhong, Y. (2014). A model approach to the estimation of peer-to-peer traffic matrices. *Parallel and Distributed Systems, IEEE Transactions on*, 25(5):1101–1111.

[Yang and De Veciana, 2004] Yang, X. and De Veciana, G. (2004). Service capacity of peer to peer networks. In *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, volume 4, pages 2242–2252. IEEE.

[Zander and Armitage, 2005] Zander, S. and Armitage, G. (2005). A traffic model for the xbox game halo 2. In *Proceedings of the international workshop on Network and operating systems support for digital audio and video*, pages 13–18. ACM.

[Zhang et al., 1996] Zhang, T., Ramakrishnan, R., and Livny, M. (1996). Birch: an efficient data clustering method for very large databases. In *ACM SIGMOD Record*, volume 25, pages 103–114. ACM.

[Zhao et al., 2009] Zhao, Q., Xu, M., and Fränti, P. (2009). Sum-of-squares based cluster validity index and significance analysis. In *Adaptive and Natural Computing Algorithms*, pages 313–322. Springer.

[Zhu et al., 2003] Zhu, C., Wang, Y., Zhang, Y., and Wu, W. (2003). Different behavioral characteristics of web traffic between wireless and wire ip network. In *Communication Technology Proceedings, 2003. ICCT 2003. International Conference on*, volume 1, pages 267–271. IEEE.

[Zink et al., 2009] Zink, M., Suh, K., Gu, Y., and Kurose, J. (2009). Characteristics of youtube network traffic at a campus network–measurements, models, and implications. *Computer Networks*, 53(4):501–514.

[Zou et al., 2013] Zou, L., Trestian, R., and Muntean, G.-M. (2013). Doas: device-oriented adaptive multimedia scheme for 3gpp lte systems. In *Personal Indoor and Mobile Radio Communications (PIMRC), 2013 IEEE 24th International Symposium on*, pages 2180–2184. IEEE.

Acrónimos

3GPP 3rd Generation Partnership Project

ADSL Asymmetric Digital Subscriber Line (Línea de Abonado Digital Asimétrica)

AIMC Asociación para la Investigación de Medios de Comunicación

AMETIC Asociación de Empresas de Electrónica, Tecnologías de la Información, Telecomunicaciones y Contenidos Digitales

ANSI American National Standards Institute (Instituto Nacional Estadounidense de Estándares)

ARIMA AutoRegressive Integrated Moving Average (Media Móvil Integrada AutoRegresiva)

ATM Asynchronous Transfer Mode (Modo de Transferencia Asíncrona)

BCSM Between Cluster Scatter Matrix

BWA Broadband Wireless Access (Acceso Inalámbric de Banda ancha)

CATV Community Antenna Television (Televisión por cable)

CIS Centro de Investigaciones Sociológicas

CMT Comisión del Mercado de las Telecomunicaciones

CMTS Cable Modem Termination System (Sistema de Terminación de Cablemódems)

CNMC Comisión Nacional de los Mercados y la Competencia

CO Central Office (Oficina Central)

CRISP-DM CRoss-Industry Standard Process for Data Mining (Proceso Estándar industrial para la Minería de Datos)

CSMA/CD Carrier Sense Multiple Access with Collision Detection (Acceso Múltiple con Escucha de Portadora y Detección de Colisiones)

DES Discrete Event Simulation (Simulación de Eventos Discretos)

DNS Domain Name System (Sistema de Nombres de Dominio)

DOCSIS Data Over Cable Service Interfaces Specification (Especificación de Interfaz para Servicios de Datos por Cable)

DSL Digital Subscriber Line (Línea de Abonado Digital)

DSLAM DSL Access Multiplexer (Multiplexor de Acceso DSL)

EGM Estudio General de Medios

ETSI European Telecommunications Standards Institute (Instituto Europeo de Normas de Telecomunicaciones)

EUROSTAT Statistical Office of the European Communities (Oficina Europea de Estadística)

FARIMA Fractional ARIMA (ARIMA Fraccional)

FBM Fractional Brownian Motion (Movimiento Browniano Fracccional)

FPS First Person Shooter (Disparos en Primer Persona)

FTP File Transfer Protocol (Protocolo de Transferencia de Archivos)

FTTH Fiber To The Home (Fibra hasta el Hogar)

GOP Group Of Pictures (Grupo De Imágenes)

GoS Grade of Service (Grado de Servicio)

GPON Gigabit-capable Passive Optical Network (Red Óptica Pasiva con Capacidad de Gigabit)

GUI Graphical User Interface

HDSL High bit rate Digital Subscriber Line (Línea de Abonado Digital de Alta velocidad binaria)

HFC Hybrid Fiber Coaxial (Híbrido de Fibra-Coaxial)

HTTP Hypertext Transfer Protocol (Protocolo de Transferencia de Hipertexto)

IAT InterArrival Time (Tiempo Entre Llegadas)

IDE Entorno de Desarrollo Integrado

INE Instituto Nacional de Estadística

IP Internet Protocol

ISP Internet Service Provider (Proveedor de Servicios de Internet)

ITU International Telecommunication Union (Unión Internacional de Telecomunicaciones)

ITU-T ITU Telecommunication Standardization Sector (Sector de Normalización de las Telecomunicaciones de la ITU)

KDD Knowledge Discovery in Databases (Descubrimiento de Conocimiento en Bases de Datos)

KDP Knowledge Discovery Process (Proceso de Descubrimiento de Conocimiento)

LAN Local Area Network

LRD Long Range Dependency (Dependencia a Largo Plazo)

LTE Long Term Evolution

MIMO Multiple-Input Multiple-Output (Múltiple Entrada Múltiple Salida)

MMOG Massively Multiplayer Online Game

MMORPG Massively Multiplayer Online Role-Playing Game

MMP Markov Modulated Process (Proceso de Markov Modulado)

MMPP Markov Modulated Process (Proceso de Poisson modulado por Markov)

MUX Multiplexor

ODN Optical Distribution Network

OECD Organización para la Cooperación y Desarrollo Económicos

OFDMA Orthogonal Frequency-Division Multiple Access (Multiplexación por División de Frecuencia Ortogonal)

OLT Optical Line Termination

ONT Optical Network Termination

ONTSI Observatorio Nacional de las Telecomunicaciones y de la Sociedad de la Información

P2P Peer-to-Peer

PAM Pulse Amplitude-Modulation (Modulación por Amplitud de Pulsos)

PON Passive Optical Network (Red Óptica Pasiva)

QoE Quality of Experience (Calidad de Experiencia)

QoS Quality of Service (Calidad de Servicio)

RTP Real-time Transport Protocol (Protocolo de Transporte de Tiempo real)

RTSP Real Time Streaming Protocol (Protocolo de Flujo en Tiempo Real)

S-CDMA Synchronous Code Division Multiple Access (Acceso Múltiple por División de Código Síncrono)

SC-FDMA Single Carrier Frequency Divison Multiple Access (Acceso Múltiple por División de Frecuencia de Portadora Única)

SCTE Society of Cable Telecommunications Engineers (Sociedad de Ingenieros de Telecomunicaciones de Cable)

SDSL Symmetric Digital Subscriber Line (Línea de Abonado Digital Ssimétrica)

SHDSL Single line High Speed Digital Subscriber Line (Línea de Abonado Digital de un solo par de alta velocidad)

SRD Short Range Dependency (Dependencia a Corto Plazo)

TAP Terminal Access Point (Punto de Acceso de Terminal)

TCP Transmission Control Protocol

TDM Time Division Multiplexing (Multiplexación por División de Tiempo)

TDMA Time Division Multiple Access (Acceso Múltiple por División de Tiempo)

TIC Tecnologías de la Información y la Comunicación

UDP User Datagram Protocol

URI Uniform Resource Identifier (Identificador Uniforme de Recursos)

URL Uniform Resource Locator (Localizador de Recursos Uniforme)

VDSL Very high bitrate Digital Subscriber Line (Línea de Abonado Digital de Muy alta velocidad binaria)

VoIP Voice over IP (Voz sobre IP)

WCSM Within Cluster Scatter Matrix

WiMAX Worldwide Interoperability for Microwave Access (Interoperabilidad Mundial para Acceso por Microondas)

WoW World of Warcraft

WWW World Wide Web